Université Paris Cité

École doctorale de Sciences Mathématiques de Paris Centre (ED 386)

Laboratoire MAP5, UMR 8145 CNRS

Computer-assisted Proofs of Reachability Analysis for Linear Control Systems under Bounded Constraints

par Ivan HASENOHR

Thèse de doctorat de Mathématiques Appliquées

Dirigée par Sébastien MARTIN, Camille POUCHOL, et Yannick PRIVAT

Soutenue publiquement le 20 Novembre 2025 devant un jury composé de :

Franck BOYER	Professeur, Université de Toulouse	Rapporteur
Maxime BREDEN	Maître de Conférences, Ecole Polytechnique	Examinateur
Sébastien MARTIN	Professeur, Université Paris Cité	Directeur
Camille POUCHOL	Maître de Conférences, Université Paris Cité	Invité
Yannick PRIVAT	Professeur, Université de Lorraine	Directeur
Aude RONDEPIERRE	Professeure, Insa Toulouse	Rapporteuse
Marcela SZOPOS	Professeure, Université Paris Cité	Présidente
Christophe ZHANG	Chercheur, INRIA Paris	Invité

Université Paris Cité
45, rue des Saints-Pères
75006 Paris

Abstract

This thesis is devoted to the reachability of linear control problems subject to bounded control constraints. Using a convex-analytic approach and supporting hyperplanes, we introduce a functional J whose nonnegativity on the state space is equivalent to the reachability of a target set. One key property of J is that it depends exclusively on control constraints and on the solution to the adjoint problem.

By means of the proposed functional, we mainly focus on proving the non-reachability of targets, which amounts to finding points at which J is negative. We introduce a computer-assisted method relying on the discretisation and minimisation of J. First, a discretised proxy J_d is minimised to find a negative point. We then bound the discretisation errors using new discretisation-error bounds with explicit constants, and rounding errors thanks to interval arithmetic. Ultimately, this ensures the original functional J's value is enclosed in a computed interval, whose negativity delivers a computer-assisted proof of the non-reachability of the chosen target.

We prove the effectiveness of the method both in the finite- and infinite-dimensional setting: in a first part, we formalise the method and develop discretisation-error bounds for a wide array of finite-dimensional systems and control constraints, and apply it to various examples, providing non-reachability proofs and lower bounds of minimal reachability times. In a second part, we extend the method to the control of linear parabolic equations, with the 1D heat equation and variants thereof as the prime example. By finely estimating constants associated to discretisation errors, we arrive at proofs of non-reachability for targets under realistic constraints.

This functional can also be used to prove positive results of reachability: we also present ongoing work aiming to compute verified under- and over-approximations of the reachable set.

Keywords: Linear control systems, bounded control constraints, reachability analysis, finite- and infinite-dimensional control systems, computer-assisted proofs, interval arithmetic.

Résumé

Cette thèse étudie l'atteignabilité de problèmes de contrôle linéaires soumis à des contraintes bornées sur le contrôle. En utilisant une approche basée sur de l'analyse convexe et des hyperplans supports, nous introduisons une fonctionnelle J dont la positivité sur l'espace d'état est équivalente à l'atteignabilité de l'ensemble cible. Cette fonctionnelle a l'avantage de dépendre uniquement des contraintes sur le contrôle et de la solution du problème adjoint.

À l'aide de cette fonctionnelle, nous nous concentrons sur les preuves de non-atteignabilité – celle-ci étant équivalente à trouver un point où J est négative. Nous introduisons une méthode assistée par ordinateur basée sur la discrétisation puis la minimisation de J. Dans un premier temps, nous minimisons une version discrétisée J_d de J, dans le but de trouver un point négatif. Ensuite, nous majorons les erreurs de discrétisation grâce à de nouvelles bornes explicites, et bornons les erreurs d'arrondis en utilisant de l'arithmétique d'intervalles. Finalement, la valeur de la fonctionnelle originelle J est incluse dans un intervalle explicite, dont la stricte négativité fournit une preuve assistée par ordinateur de non-atteignabilité de la cible choisie.

Nous démontrons l'efficacité de la méthode tant en dimension finie qu'infinie : tout d'abord, nous la formalisons dans le cadre de systèmes de dimension finie et développons des majorations explicites d'erreurs de discrétisation. Ces résultats sont appliqués à de nombreux exemples, fournissant des preuves de non-atteignabilité ainsi que des minorations de temps minimaux d'atteignabilité. Dans un second temps, nous étendons la méthode à des systèmes de contrôle paraboliques, avec l'équation de la chaleur en une dimension comme exemple type. En majorant finement les erreurs de discrétisation, nous obtenons des preuves de non-atteignabilité sous des contraintes réalistes.

La fonctionnelle J peut également être utilisée pour prouver des résultats positifs d'atteignabilité : cette thèse présente ainsi des travaux en cours sur des preuves assistées par ordinateur d'approximations tant intérieures qu'extérieures de l'ensemble atteignable.

Mots-clefs : Problèmes de contrôle linéaire, contrôle sous contraintes bornées, atteignabilité, systèmes de dimension finie et infinie, preuves assistées par ordinateur, arithmétique d'intervalles.

Contents

Remerciements

Ma thèse au MAP5 fut pour moi une source intarissable d'épanouissement, tant scientifique qu'humain. Cette époque faste et son résulat (ce manuscrit) ont été rendus possibles par une multitude de personnes que je souhaite remercier ici.

En tout premier lieu, mes trois exceptionnels directeurs de thèse, qui m'ont pris sous leurs ailes et qui m'ont encadré de façon exemplaire pendant plus de trois ans. Merci pour tous ces rendez-vous alliant qualité scientifique, digressions à n'en plus finir et délicieux jeux de mots. Votre trio fonctionnait de manière à la fois commune et complémentaire, et cet encadrement s'est révélé optimal pour moi. De manière succinte mais plus personnelle : merci Yannick pour ta perspicacité et ton expérience, pour ton extraordinaire capacité à ne jamais rechigner à un calcul ardu, et bien sûr pour ton talent d'imitateur hors pair. Christophe, merci pour tes indénombrables dessins de patates, pour ta ponctuation toujours aussi ponctuelle, et pour toutes ces discussions et découvertes musicales. Camille, un immense merci pour ton accompagnement tout au long de ces années : je n'ai cessé d'apprendre à tes côtés et j'espère que c'en est que le début ! En dernier mais non des moindres, je voudrais remercier le travailleur de l'ombre, qui n'a jamais failli d'accomplir son rôle avec humour et efficacité : merci Sébastien !

Je souhaite ensuite remercier tous ceux dont l'expertise scientifique a permis cette soutenance : tout d'abord, un immense merci à Aude Rondepierre et Franck Boyer d'avoir accepté de rapporter cette thèse, et de l'avoir fait aussi consciencieusement. Merci également, Maxime, tant pour faire partie de mon jury que pour ces discussions alliant polytopes et arithmétique d'intervalles. Un grand merci à Marcela de présider ce jury, ainsi que pour tes encouragements si réguliers ces derniers mois.

Merci également à ceux qui ont influencé le contenu de cette thèse plus directement : Kevin et Fabien, vos conseils durant mes comités de suivi ont su m'aiguiller et me rassurer à des moments charnières de ma thèse. Merci aussi à Michel Crouzeix dont les notes de cours *paboliques* ont guidé mes cheminements en théorie des opérateurs, et à Loïc Bourdin pour sa revue précise du premier article de cette thèse ainsi que pour ces intéressantes idées d'extensions.

J'en profite pour remercier ceux qui m'ont guidé au fur et à mesure de mes études : merci aux Loïcs Evanno et Laferté, ainsi que l'association MATh.en.JEANS, qui m'ont offert ma première expérience de recherche en mathématiques ; merci Stéphane Egée de m'en avoir accordé un second aperçu. Plus tard, merci aux Laurents Ducrohet et Boudin qui, successivement, ont sû me ravir avec leur humour et leur sens aigu de la pédagogie. Un grand merci ensuite à l'ensemble des professeurs du M1 MMA : ce fut un immense plaisir et honneur que de vous retrouver ensuite au MAP5. Finalement, je remercie Emmanuel Trélat pour son extraordinaire cours de M2 m'ayant fait découvrir la théorie du contrôle.

J'ai eu l'immense chance de pouvoir faire ma thèse au MAP5, qui s'est révélé, comme pour tant d'autres avant moi, un lieu où travail et plaisir savent si bien s'entremêler. Cela n'est possible que par la présence et l'investissement d'une ribambelle de personnes œuvrant à son bon fonctionnement. Tout d'abord, un grand merci à Anne et Marie-Hélène, qui m'ont dès mon arrivée accueilli et accompagné par leur gentillesse et disponibilité. Leur relève est entre temps arrivée et leur fait honneur : merci Gladys pour ton efficacité et ta présence, merci Martine pour ton aide et pour ces si nombreux bavardages. Et bien entendu, merci Antoine pour ton dévouement et ton énergie sans fin, ton humour contagieux, ton amour des échecs ou des petits mets. Merci également à Christophe Castellani, Isabelle Valero et Augustin Hangat, Arnaud Meunier pour avoir travaillé dans l'ombre aux petites choses essentielles. Merci finalement à

Amina Hariti, qui d'un peu plus loin a fourni le travail gigantesque d'incarner l'école doctorale.

Le MAP5 c'est aussi un florilège de permanents qui contribuent chacun à leur manière à une superbe ambiance. Merci Rémy pour ta si grande bonté et tes conseils algorithmiques. Marie, puisses-tu un jour céder quotidiennement à l'attrait du tarot : merci pour ton éternel sourire et ton enthousiasme. Angèle, ton dévouement et ton attention permanente envers l'ensemble des doctorants font chaud au cœur. Merci aussi à Jonathan et Félix qui se succèdent à tes tes côtés avec brio. Andrés, merci pour tes précieux éclairages sur les méandres de l'école doctorale. Un grand merci à l'esnsemble de l'équipe MAS pour tous ces exposés et journées passionnantes ; merci à Camille, Eric, Quentin et Irène, j'ai eu grand plaisir à enseigner à vos côtés. Finalement, merci à tous ceux avec qui j'ai pu partager une discussion, un rire ou une partie de tarot : notamment, merci à Emré, Paul, Fabienne, Manon, Sylvain, Nathalie, Joan, Bérénice, Georges, Raphaël, Thomas, Antoine M., Annie, Geneviève, Anne S., Assaf, Pierre-Emmanuel, Jean-Christophe et Amandine.

Autour de toute cette assemblée virevoltent un petit nuage d'éphémères : j'ai eu la change d'en faire partie pendant plus de trois ans, et me dois de leur exprimer ma reconnaissance. Tout d'abord, merci aux « anciens » qui m'ont si bien accueilli : Florian, merci pour ta fidélité au MAP5, et bien sûr d'avoir contribué à intégrer le tarot au labo ! Merci à Anton pour son bureau haut en verdure et ses initiations au yoga ; Antoine M., c'est toujours un plaisir d'échanger fromages jurassiens et savoyards ; Mariem, merci pour ta profonde bienveillance, ce fût un plaisir d'organiser la JDDM avec Sonia et toi ! Ariane, tu as le don d'illuminer les visages de ceux qui t'entourent ; merci pour tous ces moments de pur bonheur, pour cette amitié si naturelle et chaleureuse. Zoé finalement, merci pour ton sourire si contagieux et ton grand cœur. Merci également à tous ceux que je n'ai pas eu l'occasion d'apprendre à mieux connaître : Antoine S., Herb, Marie, Ousmane, Chabane, les deux Rémis, Safa, Sinda, Thaïs et Yen.

Un immense merci également aux membres du bureau 725-A1 : ce ne fût certes pas le bureau le plus peuplé, mais je me suis d'autant plus attaché à ses membres. Mehdi, maître des imprimantes, merci pour l'odeur de ton thé embaumant le bureau, pour tes si nombreux conseils et encouragements délivrés au cours d'innombrables conversations. Merci Sonia pour tes petites habitudes du bureau : chaises non rangées et siestes inopinées me font encore sourire ; merci aussi pour ton écoute et dévouement pour l'ensemble des éphémères. Votre présence quasi-continue pendant pus de deux ans fût un don inestimable. Merci Diala pour ton grand sourire et ta constante joie de vivre. Vos successeurs ont su, sans le savoir, vous faire honneur : Eyal, merci pour toutes ces distractions mutuelles quand le travail ne nous disait guère ; ton auto-dérision et la passion qui imprègnait certaines conversations m'ont beaucoup appris. Mélain, j'espère que tu t'en sortiras sans l'« ancien du bureau » ; merci pour ces nombreuses discussions – j'espère goûter un jour de ton fameux fromage béninois qui te fais oublier ton prénom ;-) Sylvain, merci pour toutes ces conversations sur des sujets divers, l'avenir te réserve de bien belles choses, crois-moi ! Merci également à Alessio et Laurent B. pour ces longues et nombreuses discussions, à Sisi pour ta bonne humeur et tes éclats de rire si francs, à César et sa bonne humeur, Raphaël (prends bien soin de Camille pour moi), au cher breton Robinson, à la si chaleureuse Oumayma, à Laurent F., à Lucca et ton impressionnante culture mathématique, ainsi qu'à Arwan, Félix, Damian, Tomasz, Ruikai, Hugo et Joël.

Les éphémères ne se résument pas au 725-A1, bien heureusement ! Là encore, commençons par les plus vieux, avec qui j'ai eu la chance de partager cette fin de thèse : Eloi, je veux t'exprimer ma plus profonde reconnaissance pour ces trois années de soutien de tous les instants, pour tous ces moments de partage hors du temps, pour ta générosité et ton don de soi permanent : merci. Merci également Xuwen, ton enthousiasme et ta candeur me fait toujours chaud au cœur – merci aussi à vous deux pour votre visite guidée de la gastronomie chinoise ! Alexander, you seem to have a gift to push people beyond their limits: thank you for doing so, as well as for all those quirky conversations, which always end up to be most entertaining, and surprisingly not completely devoid of sense. Julia, thank you for your incredible empathy and utter devotion to your friends and family: you never fail to impress me. Loïc, à nous deux nous pouvons presque concurrencer les conversations loufoques d'Alexander : merci pour ces oh combien amusants moments, pour toutes ces merveilleuses facettes de ta personnalité qu'il faut, tel un bon fromage franc-comtois (ou, paraît-il, vin bourguignon), du temps pour découvrir. Merci aussi pour ta passion des mathématiques, qui mena notamment au Lemme III.27. Keanu, c'est toujours un plaisir de redécouvrir ton étonnante connaissance des dynamiques socio-culturelles du 13ème arrondissement. Guillaume, merci pour ces longues discussions sur le futur, pour ton pragmatisme et tes nombreux encouragements. Merci également à Elisa et Léna avec qui j'ai apprécié partager des conversations au détour d'un bar ou d'un ascenceur.

Et puis bien sûr viennent tous ces jeunes à qui l'avenir sourit : Adélie, merci pour ton incontournable présence au laboratoire, pour ta sincérité, l'attention que tu portes aux autres et pour tes ébahissantes commandes de fromage. Bea, merci pour cette si grande facilité de conversation que nous partageons, pour cette profonde liberté qui émane de toi et qui teint les rires et joies parsemant nos interactions. Clémence, ce fut un immense plaisir que de découvrir ton naturel si doux et rieur, ainsi que tes talents d'organisatrice hors pair. Lucie, merci pour ton humour, ton humanité, et ton franc amour de la musique : j'adore discuter avec toi et ai hâte d'une prochaine session piano ! Merci aussi à Ariel et ses poissons migrateurs, Rayane et ses souriants bougonnements, ainsi qu'à Felix S. et Laura.

Je voudrais finalement remercier tous ceux que je n'ai pu que croiser : au nouveaux d'abord, je vous souhaite que votre thèse au MAP5 soit aussi épanouissante que la mienne. Merci à Perla pour ton amour du paon et ton incroyable bonté, merci à Pierre pour ton ardeur au tarot et tes explications modales, merci aussi à Baptiste, Timothé et Tess. Ensuite, un grand merci aux réels éphémères du laboratoire, qui ne passèrent que quelques semaines ou mois parmi nous : Thank you Bianca for your numerous visits during which you have enriched us with your warmth and kindess. Thanks as well to the inventor of the chess tactic "the Martin", it's always a great pleasure to talk with you. I hope to be able to speak with you both in German very soon! Merci Tom pour ton amitié si fidèle et sincère malgré la distance – toujours partant pour une nouvelle découverte en ta compagnie ! Merci Adrien pour ton grand humour transcendsant disputes régaliennes et voyages bavarois, j'espère encore te rendre visite dans le grand Nord. Merci finalement à tous les autres que j'aurais souhaité apprendre à mieux connaître : Félix W., Damian, Léonard, Eve, Titouan, Bernardin, Nick, Michel, Yichuan, Adéchola, Thibaut, Yassine, Angie, Nizar, Valentine, Samuel, Julie, Thomas P, Maria.

En dehors du MAP5, nombreux sont ceux qui, plus ou moins directement, ont contribué au bon déroulement de cette thèse. Un petite pensée tout d'abord à tous ceus qui l'ignorent complètement, parfois même depuis la tombe : Franz Schubert, John Powell ou Taylor Swift, John Keats ou William Blake, ou même la fruitière de la Baroche, merci pour tous ces petits plaisirs qui illuminent le quotidien !

J'ai bien sûr profité durant ma thèse du soutien de mes nombreux et fidèles amis. Tout d'abord, les seuls avec qui je peux parler tout mon saoul d'Hahn-Banach : merci à Emmanuel, Perceval, Adam et Eléana pour tous ces moments de camaraderie et de partage. Merci à mes chers amis de l'Académie du Palais Royal, avec qui j'ai passé tant de bons moments : Mehdi et Magali, Anaïs puis Charaf et Quentin. Merci à Pascal et à Quid Novi où j'ai passé d'excellentes soirées à chanter, merci Maël pour ton ambivalence musicale et scientifique, source de si belles discussions. Merci à Léo et Armel, avec qui j'aime remémorer des souvenirs de Chaptal et partager des moments de franche amitié. Merci également à mes vieux amis du BS : Albane, Luana, Matthieu, Victor. Merci aussi à Anysia pour ces longues années d'amitié, ainsi qu'à Lou et Kavya dont l'exceptionnelle amitié m'a tant soutenu à maintes reprises.

Et finalement, où en serais-je sans ma chère famille ? Vous m'avez encouragé, poussé et conseillé au quotidien, merci à vous tous : mes grand-parents tout d'abord, merci Grand-Mère et Grand-Père, merci Mayayé, vous êtes fabuleux, chacun à votre manière. Merci infiniment Anne-Sophie, David, Jules, Lucie, Anaëlle pour ces si nombreux accueils et cet incroyable don de soi dont vous faites toujours preuve. Merci à la famille plus éloignée que je revois toujours avec plaisir : mes oncles et tantes et cousins d'à travers le monde. Merci à Milou et Pierre, mes deux frères géniaux avec qui j'ai bien hâte d'aller explorer le continent sud-américain, et finalement un immense merci à mes deux parents, qui ont simplement tout rendu possible d'une multitude indénombrable de façons. Merci, merci à tous.

0

Introduction

Contents

0.1 Introduction to control theory

Control theory is a field of applied mathematics concerned with the study of dynamical systems which can be influenced by an exogenous action. As such, it has wide applications ranging from mechanical devices (aeronautics, robotics) to chemistry as well as quantum physics, economy, biology... A *control system* is usually a system of (partial) differential equations in which at least one parameter – the *control* – can be chosen by the user. The choice of this control will then determine the *trajectory* of the system.

For instance, a car's motion can be described using three parameters: its position, direction and speed – those define the *state* of the system. Since the driver can influence those parameters, making it accelerate, slow down or change directions, a car is a control system for which the pedals and the steering wheel form the control. The choice and effect of the control induced by the driver are tempered by *state* or *control constraints*: state constraints describe exogenous restrictions, such as traffic laws – speed limits, safety measures. On the other hand, control constraints restrain the control itself: mechanical properties such as engine and brake power, steering wheel maximum angle or gas tank capacity are a few examples.

When studying this system, a few questions are of major interest: firstly whether the car can be driven from one configuration (position, direction, speed) to another in a given amount of time – this is called a *reachability problem*. Obviously, this heavily depends on many parameters, such as the state and control constraints, the available time and the relative position of both initial position and target. For instance, while a standard car might be steered from Paris to Orleans within two hours, Lyon is unreachable in the same timelength. Similarly, a car without reverse gear will not be able to parallel park.

In control theory, a related problem is *controllability*: a system is called *controllable* if it can be steered from any initial position to any target in a given time – usually, without considering any state or control constraints. This property is important for it provides an accurate description of the inner dynamics of the system and its control.

Once the reachability problem has been solved, a next natural question is *how* to do it: consider a car going from Paris to Bordeaux. Many options are available: the car could go straight from Paris to Bordeaux, or the driver might decide to take a nap at Poitiers – or even Vienna, if he fancies it. Since any feasible trajectory ending with a stop at Bordeaux is acceptable, however peculiar it is, one can wonder what the optimal trajectory, and thus what the matching *optimal control*, is.

Of course, "optimal" can have many different meanings, depending on the main criterion under consideration. Maybe the first that comes to mind is *time-optimality*: what is the control and controlled trajectory minimising the transfer time? In our example, it is the logical one: start from Paris, go as fast and as straight as possible without any stops and brake at the last minute to stop in Bordeaux. One can see how this brutal strategy might not be optimal if considering other criteria such as safety or energy efficiency.

Another very commonly studied problem is that of *energy-optimality*. When driving a car, two main factors influence gas consumption: speed (the faster one drives, the more energy is consumed), and acceleration (changing speed increases overall consumption). Intuitively, it follows that driving as slowly and steadily as possible over the shortest route seems optimal energy-wise. However, the optimal control can vary depending on the specific constraints and available transfer time; relaxing these may lead to different optimal strategies.

Of couse, in practice, multiple criteria have to be jointly considered: time, energy, but also for example cost and safety. The desired combination of these criteria and constraints provide an appropriate cost functional that control theory aims at minimising.

To conclude this brief introduction, these are the main objectives of control theory: given a dynamical system, what targets are reachable? And if they are, what is an optimal control to reach them – given an appropriate criterion? In this thesis, we are mostly concerned with the question of reachability, even though optimal control will play its part. We study general abstract linear systems, but apply our results and methodology to a few examples, such as models of a streetcar and of a space capsule in Chapter II and diffusion of heat along a one-dimensional bar in Chapter III.

0.2 Reachability analysis of linear autonomous control systems

In this thesis, we consider linear autonomous (time independent) control systems. Consider X a Hilbert space, referred to as the *state space*, and U another Hilbert space – the *control space*. Both are endowed with the norms and inner products $\|\cdot\|_X$ and $\langle\cdot,\cdot\rangle_X$ (respectively $\|\cdot\|_U$ and $\langle\cdot,\cdot\rangle_U$). We shall also use the space $E = L^2(0,T;U)$ endowed with its standard norm and inner product $\|\cdot\|_E$ and $\langle\cdot,\cdot\rangle_E$. Where no confusion is possible, the subscripts might be dropped for readability purposes. Both spaces can be finite-dimensional, typically for the control of Ordinary Differential Equations (ODEs), or infinite-dimensional when considering Partial Differential Equations (PDEs).

We introduce the operators $A : \mathcal{D}(A) \subset X \to X$ and $B : U \to X$. B will furthermore be assumed bounded, and A will generate a C_0-semigroup $(S_t)_{t\geq 0}$. We will consider the following linear control problem:

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{for a.e. } t \in [0,T], \\ y(0) = y_0, \\ u(t) \in \mathcal{U} & \text{for a.e. } t \in [0,T], \end{cases} \tag{$\mathcal{S}$}$$

where $T > 0$ and $\mathcal{U} \subseteq U$ will embody time-independent constraints on the distributed control. We will furthermore denote

$$E_{\mathcal{U}} = \{u \in E,\ u(t) \in \mathcal{U} \text{ for a.e. } t \in [0,T]\} \tag{0.2.1}$$

the set of $\mathcal{U}$-admissible controls. Solutions of ($\mathcal{S}$) can be characterised using Duhamel's formula:

$$y(\cdot\,; y_0, u) : t \mapsto S_t y_0 + L_t u, \tag{0.2.2}$$

where the input-operators $(L_t)_{t\geq 0}$ are defined as

$$\forall\, t \geq 0, \qquad L_t : \begin{cases} L^2(0,t;U) & \to X \\ u & \mapsto \int_0^t S_{t-s}Bu(s)\,\mathrm{d}s. \end{cases} \tag{0.2.3}$$

Notice that this context could be generalised, for example by considering time-dependent control or state constraints, time-dependent operators A and B or by adding a drift term in $\mathcal{S}$: although these can complicate considerably the analysis, they do not change the following definitions of this introduction.

Let us now define the notions introduced in Section 0.1, starting with reachability analysis – existence of a control – and then briefly introducing optimal control.

Reachability analysis. This problem consists in trying to quantify the influence an external user has on the system: by exploiting the control variables as much as possible, what configurations of the system can be achieved? Let us introduce two essential definitions: reachability and controllability.

Definition 0.1. Given an initial condition $y_0 \in X$, a target $y_f \in X$, a time horizon $T > 0$, and control constraints $\mathcal{U} \subset U$, we say that y_f is $\mathcal{U}$-reachable from y_0 in time T if there exists $u \in E_{\mathcal{U}}$ and

$$y(T\,; y_0, u) = y_f. \tag{0.2.4}$$

The set of $\mathcal{U}$-reachable states from y_0 in time T is called the reachable set (from y_0 in time T under constraints $\mathcal{U}$), and is denoted

$$\mathcal{R}_{T,\mathcal{U}}(y_0) = \{y_f \in X, \quad y_f \text{ is } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T\}\,. \tag{0.2.5}$$

The union over time of the reachable sets, $\bigcup_{T\geq 0} \mathcal{R}_{T,\mathcal{U}}(y_0)$, is sometimes called the reachable tube. In the unconstrained case $\mathcal{U} = U$, we shall drop the index $\mathcal{U}$.
See Figure 0.1 for an illustration of the reachable sets and tube.

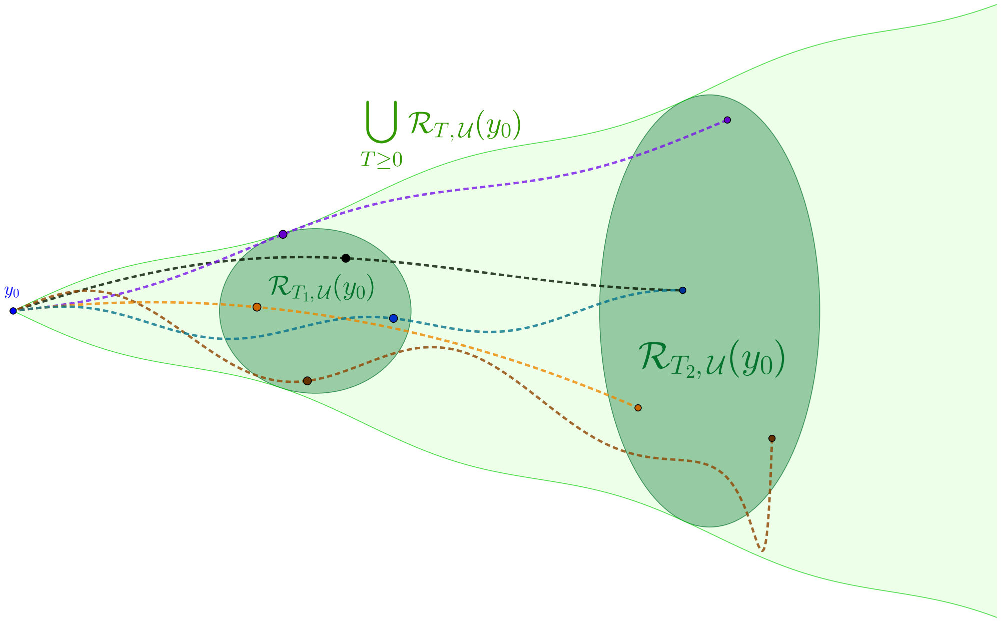

$\bigcup_{T\geq 0} \mathcal{R}_{T,\mathcal{U}}(y_0)$
y_0
$\mathcal{R}_{T_1,\mathcal{U}}(y_0)$
$\mathcal{R}_{T_2,\mathcal{U}}(y_0)$

Figure 0.1: Reachable sets, reachable tube and some controlled trajectories.

Notice that, by linearity, we have the following characterisation of the reachable set:

$$\mathcal{R}_{T,\mathcal{U}}(y_0) = \{S_T y_0\} + L_T E_{\mathcal{U}}, \tag{0.2.6}$$

and thus the necessary and sufficient condition of $\mathcal{U}$-reachability

$$y_f \text{ is } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T \qquad \Longleftrightarrow \qquad y_f - S_T y_0 \in L_T E_{\mathcal{U}}. \tag{0.2.7}$$

A weaker notion of reachability would be the one of *approximate reachability*: a target y_f is approximately reachable if it lies in the closure of the reachable set. In this thesis, since the control constraints considered are convex, bounded and closed, the reachable set will be closed and thus both notions are equivalent.

Another similar problem commonly studied – and surprisingly enough often easier to tackle – is *controllability*:

Definition 0.2. The control system $(\mathcal{S})$ is said to be *$\mathcal{U}$-controllable* in time $T > 0$ if any target y_f is $\mathcal{U}$-reachable from any initial condition y_0 in time T, that is, if

$$\forall\, y_0 \in X, \qquad \mathcal{R}_{T,\mathcal{U}}(y_0) = X. \tag{0.2.8}$$

If no control constraints are considered ($\mathcal{U} = U$), controllability boils down to linear algebra: the surjectivity of the input operator L_T. This property is well understood in finite dimension, with simple characterisations such as the Kalman rank condition. In infinite dimension, it is much more difficult to tackle, and usually involves proving elusive *observability inequalities*: the surjectivity of L_T is equivalent to the following quantitative injectivity property of L_T^* (for instance, see [J-M07])

$$\exists\, C_T > 0,\ \forall\, p_f \in X, \qquad \|L_T^* p_f\|_E \geq C_T \|p_f\|_X. \tag{0.2.9}$$

All these characterisations are very powerful tools when considering unconstrained control, but they do not apply when considering constraints on the control. In particular, bounded control constraints preclude both controllability and its weaker counterpart, *approximate controllability.*

Reachability or controllability properties can as well be proved using *optimal control theory* – which is much more resilient to the addition of control constraints.

Optimal control. Assume now that, beyond proving the reachability of a given target, one also seeks to minimise a cost functional J, which may depend on the control, the trajectory, and the final time. Assuming it only depends on the control, the minimisation problem can then be written as

$$\inf_{\substack{u\in E_{\mathcal{U}} \\ y(T;\,y_0,u)=y_f}} C(u). \tag{0.2.10}$$

Many techniques can be used to prove the existence of minimisers of this functional, typically involving continuity and coercivity conditions, and in many contexts, convex analysis – Section I.2 introduces many important results coming from this field. Notice that if the existence of such minimisers requires the $\mathcal{U}$-reachability of y_f from y_0 in time T, the converse is also true – we will take advantage of this fact in this thesis. Furthermore, considering such a functional can often provide additional information about the minimiser, as illustrated by the classical example of the Hilbert Uniqueness Method (HUM).

The Hilbert Uniqueness Method consists in considering as cost the simple squared norm $C = \frac{1}{2}\|\cdot\|_E^2$ on E. Then, considering the unconstrained case $E_{\mathcal{U}} = E$, the reachability constraint can be reformulated using indicator functions, leading to

$$\inf_{\substack{u\in E \\ y(T;\,y_0,u)=y_f}} \frac{1}{2}\|u\|_E^2 = \inf_{u\in E} \frac{1}{2}\|u\|_E^2 + \delta_{\{y_f - S_T y_0\}}(L_T u) \tag{$\mathcal{P}$}$$

where

$$\delta_{\{y_f - S_T y_0\}}(L_T u) = \begin{cases} 0 & \text{if } L_T u = y_f - S_T y_0 \\ +\infty & \text{otherwise.} \end{cases} \tag{0.2.11}$$

This new functional is convex with respect to u, and one can derive that a necessary and sufficient condition for a control $u^\star \in E$ to minimise C is that $L_T u^\star = y_f - S_T y_0$ and $u^\star$ lies in the range of L_T^*. A more tractable characterisation can be obtained using Fenchel-Rockafellar duality. Consider the minimisation problem

$$\inf_{p_f \in X} \frac{1}{2}\|L_T^* p_f\|_E^2 - \langle y_f - S_T y_0, p_f \rangle. \tag{$\mathcal{D}$}$$

This so-called *dual problem* exhibits strong duality with the *primal problem* ($\mathcal{P}$), in the sense that

$$\inf_{u\in E} C(u) + \delta_{\{y_f - S_T y_0\}}(L_T u) = - \inf_{p_f \in X} \frac{1}{2}\|L_T^* p_f\|_E^2 - \langle y_f - S_T y_0, p_f \rangle\,, \tag{0.2.12}$$

and the infimum of the primal problem is attained if finite.

The existence of a minimiser for the dual problem follows from the observability inequality (0.2.9) (and so independently of y_0 and y_f); furthermore, the minimisers $(u^\star, p_f^\star)$ of the primal and dual problems satisfy $u^\star = L_T^* p_f^\star$. Thus, casting the controllability problem as an optimisation problem can lead to a constructive control.

In this thesis, we will use constrained optimal control problems with appropriate cost functions and study their Fenchel-dual problem in order to prove reachability or non-reachability results.

0.3 Review of the literature

In this section, we shall make a review of the literature of the various fields that contributed to this thesis: control theory mainly, but also control engineering, as well as computer-assisted proofs in a more general way. This section is organised as follows:

First, we will focus on introducing the main controllability results in finite-dimensional control theory: classical control results but also more recent results obtained in the control engineering community. Special attention will be paid to reachability analysis when constraints are applied on the control.

Next we review the main results of controllability of linear partial differential equations, especially of parabolic equations. Again, constraints, whether bounded or unbounded, will be given special consideration.

Finally, we will give a quick overview of computer-assisted proofs, centred on rigorous numerics for Partial Differential Equations.

0.3.1 Finite-dimensional linear control systems

In this section, we shall consider a finite dimensional linear control system of the form

$$\begin{cases} y'(t) = A(t)y(t) + B(t)u(t) \quad \forall\, t \in [0,T] \\ y(0) = y_0 \in \mathbb{R}^n, \end{cases} \tag{0.3.1}$$

where $A \in L^2([0,T];\mathcal{L}(\mathbb{R}^n))$ and $B \in L^2([0,T];\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n))$ are two matrices, and the control $u \in L^2(0,T;\mathbb{R}^m)$ may be constrained to take values inside $\mathcal{U} \subset \mathbb{R}^m$ for almost every time $t \in [0,T]$. We first present classical result of control theory to characterise controllability in the constraint-free case or with unbounded constraints. Then, we will present various methods allowing for the approximation of the reachable set when considering bounded constraints on the control.

0.3.1.a Controllability

Unconstrained control systems. We will first present very standard controllability results that can be found in every control theory textbook – see for example [LM86; J-M07; TW09].

The study of controllability – the property allowing one to steer the system from any starting point to any target configuration – goes back to the early 60's with Kalman's seminal works, including the well-known Kalman rank condition: this necessary and sufficient condition links controllability of an autonomous linear control system to the rank condition

$$\mathrm{rk}(B; AB; \ldots; A^{n-1}B) = n. \tag{0.3.2}$$

This condition is independent from the final time. It is of course a very specific feature of unconstrained controllability. For non-autonomous systems, one can define the Gramian matrix

$$\mathcal{G} = \int_0^T R(T,\tau)B(\tau)B(\tau)^* R(T,\tau)^* \,\mathrm{d}\tau, \tag{0.3.3}$$

where $R : [0,T]^2 \to \mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)$ satisfies, for $t_2 \in [0,T]$,

$$\forall\, t_2 \in [0,T],\ R(\cdot,t_2) \text{is the solution of the Cauchy problem } \begin{cases} \partial_t M = A(t)M \\ M(t_2) = \mathrm{Id}\,. \end{cases} \tag{0.3.4}$$

The invertibility of $\mathcal{G}$ constitutes a convenient necessary and sufficient condition for controllability. Both the Kalman rank condition and the Gramian characterisation are very classical controllability criteria that can be found in the aforementioned textbooks. However, they are not constructive. Another method called the Hilbert Uniqueness Method (HUM) was introduced by Jacques-Louis Lions in [Lio88; Lio92]. This method consists of studying the adjoint equation

$$\begin{cases} p'(t) = -A^* p(t) \quad \forall\, t \in [0,T] \\ p(T) = p_f \end{cases} \tag{0.3.5}$$

and to notice that every L^2-optimal control can be written as $t \mapsto B^*p(T-t)$ for some final condition $p_f \in \mathbb{R}^n$. This method, which is closely related to the duality between controllability and observability, is perfectly generalisable to infinite-dimensional systems.

Control systems with unbounded constraints. From now on, we shall consider control constraints $\mathcal{U} \subset \mathbb{R}^m$. Many such constraints have been studied: the simplest case is considering restriction of the control on a linear subspace of $\mathbb{R}^m$, but using a change of variable, this is equivalent to considering a smaller control space with a modified operator B. A very common type of unbounded constraints is requiring the control to lie inside a linear cone – one example of this is nonnegative control constraints. Some powerful geometric tools have been developed to tackle such problems, both in this linear context and in various nonlinear configurations: we refer to classical references for more details [SL12].

In the literature, constraints on the state have also been considered. In that context, controllability – restricted by state constraints – can be subject to minimal times (see for example [Kra08; LTZ17; LTZ18; BDM21; LTZ21]); in this thesis we will however focus solely on control constraints.

0.3.1.b Reachability analysis under bounded constraints

When considering control systems with bounded constraints on the control, controllability is impossible. Therefore, the literature focuses on the reachable set – the set of all targets reachable from a given initial condition in finite time, or given a final time $T \geq 0$ before or at this time T. Two categories of results exist: characterisations of the reachable set and what it contains or not, or methods to approximate it. Those two approaches have been investigated separately by two communities: control theorists and control engineers, and their combination provides an interesting overview of reachable set approximations. We shall explore both communities separately.

Characterisations of the reachable set. Bounded constraints are inherent in realistic control problems, and therefore have been considered for a long time: however, the high number of different constraints and their specificities significantly complexify their study. An early – and essential – result in the field of optimal control is the well known Pontryagin Maximum Principle, which provides necessary optimality conditions for a state-control pair to be optimal with respect to a given cost functional and control constraints (see, e.g. [SL12; Tré23]). However, one has to wait to the 70's to obtain the main results treating controllability under bounded control constraints: although controllability on a bounded time interval is impossible, one can obtain controllability-like results by dropping constraints, such as studying null-controllability under compact constraints lying near 0 [SY71; Bra72; SV86; FHL92; HLQ02], or constrained controllability without fixed final time [SB80; Vel88].

Another major characterisation of the reachable set lies in the celebrated bang-bang theorem (see, e.g., [SL12]): if the control constraint set $\mathcal{U}$ is nonempty and compact, then the reachable set itself is convex and compact, and is furthermore equal to the set of reachable points using only controls valued in the extreme points of $\mathcal{U}$.

Approximations of the reachable set. In the control engineering community, numerous methods have been developed to numerically approximate the *reachable set* (targets reachable at time T) and *reachable tube* (targets reachable before time T). Approximating the reachable set can be done for multiple purposes: the most obvious being to determine the existence of a control achieving the target, in order to then construct a control steering the system towards it. On the other hand, the reachable set is heavily studied for safety purposes, to guarantee that the system will not enter a dangerous zone. These two objectives have spawned many techniques to tackle one or both purposes and create under-approximations (included in the reachable set, also called inner-approximations) and over-approximations (which contain it, also called outer-approximations) of the reachable set. Most of the literature has focused on the computation of over-approximations of the reachable set, mainly for its immediate application to the safety of various systems – e.g. automated cars and airspace management – as well as their relatively simple formulations. In comparison, computations of under-approximations are known to be trickier, but have received increased attention in the past few years, in light of safety considerations as well as applications to robotics and medicine [XSE16]. The approximation of reachable sets is connected to many fields from which results may be used: for example, differential inclusions, differential games, or the study of regions of attraction of differential equations.

We shall now summarise the main approaches studied in the control engineering field: for details, see the recent surveys on set propagation [AFG21], Hamilton-Jacobi equation methods and barrier certificates [Wab+23], or the literature review in [WKA24] for under-approximations.

One of the most studied methods relies on the reformulation of reachable sets as a zero sublevel set of a viscosity solution to a Hamilton-Jacobi partial differential equation (HJE) [MBT05; Fis+15; CT18; Wab+23]. In particular, this method allows the study of *backwards reachable sets* (BRSs) and *backwards reachable tubes* (BRTs) for safety purposes: BRSs describe the initial conditions that could be steered to a danger zone at final time T, while BRTs include all initial conditions that could pass through a danger zone before time T.

These zero sublevel sets are then typically approximated by discretising the state space with a grid and computing various values at each point of the grid – a very precise description of the reachable sets, yet heavily affected by the curse of dimensionality. These methods have yielded remarkable results in the study of over-approximations of (backward) reachable sets and tubes for both linear and nonlinear control systems, and have recently been studied as well to provide under-approximations [XFZ19].

Another extensively studied method relies on Control Barrier Functions (CBF) [PJ04; Ame+16; Gur+18; KA18; KBH18; Wab+23]: these are functions designed so that their sign cannot change under the dynamics of the controlled system. Many techniques – often based on Lyapunov stabilisation methods – have been elaborated to design CBFs, which in turn allow for the computation of sets invariant by the system's behaviour, controls or disturbances. These invariant sets can then characterise over-approximations of the reachable set. In contrast to the previous method, this one scales better to higher dimensions, provided one first manages to synthesise an appropriate function.

Some methods endeavour to combine the best of both worlds, see [Wab+23] for abundant details and references – which also reviews the literature of predictive control and data-driven methods that we shall not detail here. A method similar to barrier functions, developed in [KHJ13], could also allow for the characterisation of an under-approximation of backward reachable sets.

Last but not least, set propagation [AFG21] allows for both under- and over-approximations. This general method uses set-based arithmetic to iterate under- and over-approximations of the reachable sets on a discrete time grid. The efficiency of this method is highly dependent on the type of set representation chosen: many have been studied, ranging from convex sets for linear dynamics (e.g. support functions [PN71; Vel92; GL08; LG09; LG10; Bai+07], polytopes [HK06; ASB10], zonotopes [Le 09; Imm15] or ellipsoids [KV02a; KV02b]) to various non-convex sets (e.g. unions of convex sets, polynomial zonotopes or star sets [AFG21]), each with their own literature, advantages and drawbacks. Set propagation is a very efficient way to approximate the reachable set, but is subject to the wrapping effect (exponential build-up of approximation errors). Numerous algorithms have thus been developed to circumvent it in each specific case and to avoid unnecessary over-approximations.

Set propagation has also proved to be an efficient way to compute under-approximations of the reachable set: the literature in this regard is quite rich, from the use of support functions to deduce polytopal under-approximations from over-approximations [Var00; Le 09; GLM06], ellipsoids [KV02b], polytope iterations [XSE16], cleverly chosen intervals [GP17], zonotopes [Ser20], and combinations of polytopes and zonotopes [WKA24].

Finally, let us mention that regarding approximations of reachable sets, very few papers have considered the influence of round-off errors on the computation of under- and over-approximations: to our knowledge, only a handful of authors led by F. Immler [Imm15] have considered the possibility of adding rigorous numerics to provide truly computer-assisted proofs of (non-)reachability. For details about rigorous numerics, see Section 0.3.3 below.

0.3.2 Infinite-dimensional linear control systems

The study of infinite-dimensional systems is notoriously more difficult than that of finite-dimensional systems, and the immense diversity of systems contributes to the very abundant literature. In this section, we shall focus on linear control systems, rapidly introduce classical methods to obtain the controllability of infinite-dimensional systems and then focus on the controllability of parabolic partial differential equations. For now, consider the general unconstrained controlled linear equation

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{for a.e. } t \in [0, T] \\ y(0) = y_0 \in X & \\ u(t) \in U & \text{for a.e. } t \in [0, T]. \end{cases} \tag{0.3.6}$$

First, recall the basic definitions: a system is said to *exactly controllable* if for any initial condition, the reachable set covers the whole state space. It is said to be *approximately controllable* if the reachable set is only dense in the state space, and *null-controllable* if 0 is reachable from any initial condition. Those three controllability definitions are intrinsically linked together: for example, in finite dimension, approximate and exact controllability are equivalent. For the wave equation, null-controllability and exact controllability are equivalent as well.

Another essential property of linear control systems is *observability.* Observability is a multi-faceted condition on solutions of the backward adjoint equation defined as

$$\begin{cases} p'(t) = -A^*p(t) & \text{for a.e. } t \in [0, T] \\ p(T) = p_f. & \end{cases} \tag{0.3.7}$$

Namely, observability is characterised by so-called observability inequalities that typically take the form

$$\forall p_f \in X, \quad \|t \mapsto B^*p(t)\|_1 \geq C_T \|p_f\|_2 \tag{0.3.8}$$

for given norms $\|\cdot\|_1$ and $\|\cdot\|_2$ respectively on $\mathcal{L}(U, X)$ and X – these norms may vary in order to obtain different versions of observability. Without going into details, observability is a dual and often equivalent notion of controllability, and has been one of the main tools to obtain controllability results in the past forty years. For details, see one of the many textbooks written on the subject [MZ04; J-M07; TW09; Boy22; Tré23].

Observability inequalities, often obtained through Carleman estimates, have yielded numerous core controllability results, such as the null-controllability of the heat equation [LR94] or the controllability of the wave equation under the Geometric Control Condition [BLR92]. Many new results have been discovered considering the reachable set of the heat equation, which is the focus of Chapter III: either for boundary control [DE18; HKT20; ELT22; KNT22] or for distributed control [CR22]. Let us mention a few recent surveys [Egi+20; CM24] and classical textbooks that treat this subject [MZ04; J-M07; TW09; Boy22; Tré23].

The controllability of partial differential equations has also been studied when constraints are applied on the control, though less so than in finite dimension. One classical example of constraint is a unilateral or conic constraint, which can reduce the reachable set and induce minimal times of reachability [Kla96; Res05; LTZ17; PZ18; PZ19; Ant+24]. Sparsity or non-convex constraints have also been studied [Zua10; PTZ24] as well as projection constraints [Erv20]. Even though they are not covered in this thesis, state constraints have also been considered in the literature; see for example [LTZ17; LM21].

In this thesis, we consider pointwise bounded control constraints. Those constraints naturally forbid controllability and small time reachability, and if the constraints are (pointwise) (weakly) compact, then the reachable set is weakly compact, and thus closed – in that context, approximate and exact reachability are equivalent. If B is furthermore bounded, the reachable set has also be proved to be convex: see [LY12, Chapter 7, Section 5.1]. Weakly compact control constraints have been considered since the 80's [Ahm85], but the literature has exploded in the last few years, either considering general pointwise constraint sets [Ber14; Ber19; Ber20], L^∞ constraints [Wan08; PTZ19] or even L^1 [CK22].

0.3.3 Computer-assisted proofs for PDEs

In this thesis, we aim at developing computer-assisted proofs for control theory. Computer-assisted proofs have been in development since the 1960s, using various techniques. Perhaps the first method employed was using computers to *prove by exhaustion*: when a large but finite number of cases are considered, one can make a computer check each one – this is the case of the proof of the famous four-colour theorem in graph theory. Another method relies on proof assistants: these are usually softwares with builtin logic that can check whether a proof's validity – Rocq (previously named Coq) and Lean are two major examples, but others exist. These proof assistants can be equipped with *automatic theorem provers*, which can search for formal proofs using various heuristics or AI-powered methods. Related methods of computer-assisted proofs rely on *symbolic computation* to provide proofs of cumbersome computations – the library SymPy in Python or the online computational intelligence Wolfram Alpha are two examples.

Another field of computer-assisted proofs is termed *rigorous numerics* or *validated numerics*. This field aims at computing and solving equations numerically while checking their correctness along the way by bounding the potential rounding errors due to floating-point arithmetic (see I.4 for details). This is the general method we use in this thesis.

Rigorous numerics have been developing since the 1980s, and many libraries now offer access to set-based arithmetic guaranteeing freedom from rounding errors – in this thesis, we used IntLAB [Rum99], developed by Prof. Siegfried Rump. Validated numerics have already been applied to the study of ordinary and partial differential equations: one classical method is to numerically certify the contractiveness of a given functional, and use a fixed-point theorem to obtain existence and uniqueness of solutions [Day+04; Bal+18; Ber+21; BBS24]. Other methods using for example rigorous spectral bounds have also been explored, see the surveys [Góm19; Kap+21], or [NPW19] for numerous rigorous numerics on elliptic equations. Another use of rigorous numerics related to our approach is the already mentioned computation of over-approximations of reachable sets [Imm15]. In Chapters II and III, we use interval arithmetic to guarantee bounds on the solution of a differential equation, which are then used as a criterion for the non-reachability of a target.

0.4 Summary of the thesis

We shall now summarise the content of the thesis. For a French translation of this summary, see Section 0.5.

In this thesis we are concerned with the reachability of linear autonomous control systems. Considering an operator A generating a strongly continuous semigroup $(S_t)_{t\geq 0}$, and a continuous operator B, we study control systems of the following form

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{for a.e. } t \in [0,T], \\ y(0) = y_0 \in \mathcal{Y}_0, & \\ u(t) \in \mathcal{U} & \text{for a.e. } t \in [0,T], \end{cases} \tag{$\mathcal{S}$}$$

where $\mathcal{U}$ denotes time-independent control constraints, $\mathcal{Y}_0 \subset X$ a set of initial conditions and

$\mathcal{Y}_f \subset X$ a set of targets. Typically, the following assumptions will be made:

$$\begin{cases} \mathcal{U} \subset U & \text{is nonempty, closed, convex and bounded,} \\ \mathcal{Y}_f \subset X & \text{is nonempty, closed and convex,} \\ \mathcal{Y}_0 \subset X & \text{is nonempty, closed, convex and bounded.} \end{cases} \tag{0.4.1}$$

Slightly weaker assumptions will also be considered (see (0.4.18)).

We consider the following non-standard set reachability definition, which coincides when considering single targets or initial sets: we say that $\mathcal{Y}_f$ is $\mathcal{U}$-reachable from $\mathcal{Y}_0$ in time T if there exists $(y_0, y_f) \in \mathcal{Y}_0 \times \mathcal{Y}_f$ such that y_f is $\mathcal{U}$-reachable from y_0 in time T.

The fundamental goal of this thesis is answering the following question:

> Given a linear control system subject to bounded control constraints $\mathcal{U}$, an initial set $\mathcal{Y}_0$, a time T and a target set $\mathcal{Y}_f$, is $\mathcal{Y}_f$ $\mathcal{U}$-reachable from $\mathcal{Y}_0$ in time T?

We will only consider controls in open-loop, and will mainly propose ways of obtaining negative answers to this question – which also precludes closed-loop reachability.

The summary of this thesis is divided in the following way:

- Subsections 0.4.1 and 0.4.2 focus on a convex-analytic problem, studying a separating functional J and linking its evaluation to the existence of a point in the intersection of two convex sets
- Subsections 0.4.3 and 0.4.4 introduce a methodology that applies this functional to the reachability analysis of linear control systems
- Subsections 0.4.5, 0.4.6 and 0.4.7 summarise the main results obtained in this thesis, both leading to and using the aforementioned methodology. Those sections are each dedicated to the three main chapters of this thesis (Chapters II, III and IV)
- Finally, Subsection 0.4.8 presents an outlook on future work in the line of this thesis, and outlines the plan of the thesis.

0.4.1 Convex analysis

We will first introduce the problem from a convex analytic point of view – see Section I.2 for details about this framework. Consider a Hilbert space X and two convex subsets $\mathcal{R}, \mathcal{Y} \subset X$. Our focus will be on methods designed to resolve the question:

$$\text{Is } \mathcal{R} \cap \mathcal{Y} \text{ empty?} \tag{0.4.2}$$

This is a classical question of convex analysis, which can be reformulated using a convex optimisation problem:

$$\pi := \inf_{x \in X} \delta_{\mathcal{R}}(x) + \delta_{\mathcal{Y}}(x), \tag{$\mathcal{P}$}$$

where δ_C is the convex indicator function of C, valued 0 on C and $+\infty$ elsewhere. Considering this somewhat crude optimisation problem, we have the equivalence

$$\mathcal{R} \cap \mathcal{Y} \neq \emptyset \quad \Longleftrightarrow \quad \pi = 0. \tag{0.4.3}$$

Of course, this optimisation problem could be regularised in an infinite number of ways while preserving that property, but its current form will suffice here.

Another approach to tackling (0.4.2) is using a separation argument in the style of the Hahn-Banach theorem: assuming that

$$\begin{cases} \mathcal{R} \text{ and } \mathcal{Y} \text{ are nonempty, closed and convex} \\ \mathcal{R} \text{ or } \mathcal{Y} \text{ is bounded,} \end{cases} \tag{0.4.4}$$

we have the following equivalence

$$\mathcal{R} \cap \mathcal{Y} = \emptyset \quad \iff \quad \exists\, p \in X,\ \sup_{x \in \mathcal{R}} \langle p, x \rangle < \inf_{y \in \mathcal{Y}} \langle p, y \rangle. \tag{0.4.5}$$

Let us denote $\sigma_C : p \mapsto \sup_{x \in C} \langle p, x \rangle$ the support function of a closed convex C, and

$$J : \begin{cases} X & \to \mathbb{R} \\ p & \mapsto \sigma_{\mathcal{R}}(p) + \sigma_{\mathcal{Y}}(-p). \end{cases} \tag{0.4.6}$$

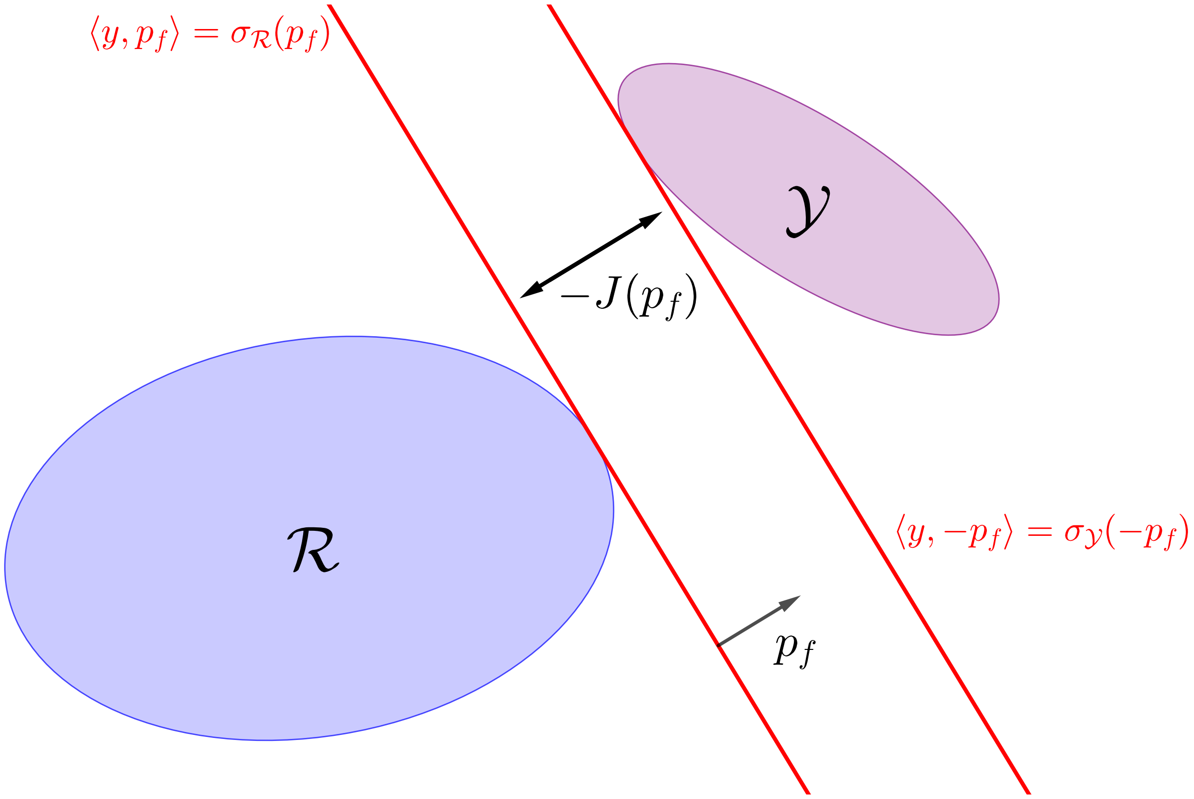

$\langle y, p_f \rangle = \sigma_{\mathcal{R}}(p_f)$
$\mathcal{Y}$
$-J(p_f)$
$\mathcal{R}$
$\langle y, -p_f \rangle = \sigma_{\mathcal{Y}}(-p_f)$
p_f

Figure 0.2: Separation of $\mathcal{R}$ and $\mathcal{Y}$ using J.

Since J quantifies the separation between $\mathcal{R}$ and $\mathcal{Y}$ (see Figure 0.2 for an illustration), it will sometimes be referred to as the *separating functional*. Furthermore J is 1-homogeneous, and thus the following minimisation problem

$$d := \inf_{p \in X} J(p) \tag{$\mathcal{D}$}$$

has only two potential values: $-\infty$ or 0. Nonetheless, under the hypotheses (0.4.4) the equivalence seen in (0.4.3) holds here as well:

$$\mathcal{R} \cap \mathcal{Y} \neq \emptyset \quad \iff \quad d = 0. \tag{0.4.7}$$

The two aforementioned optimisation problems – the *primal problem* ($\mathcal{P}$) and the *dual problem* ($\mathcal{D}$) – are in fact linked using Fenchel duality, which guarantees the *weak duality*, in the sense that

$$\pi \geq -d. \tag{0.4.8}$$

Moreover, under the assumptions (0.4.4), *strong duality* holds:

$$\pi = -d. \tag{0.4.9}$$

This is a direct consequence of the Fenchel-Rockafellar theorem [Roc67] (see Theorem I.24). In this thesis we will make use of this duality, and especially of the separation functional, developing computer-assisted proofs to compute verified approximations of it and exploiting these to answer reachability problems. Let us now, still in this convex analysis framework, provide some details about rigorously evaluating J.

0.4.2 Rigorous evaluation of the separating functional

We will henceforth assume that both $\mathcal{R}$ and $\mathcal{Y}$ are convex, closed and nonempty, and that $\mathcal{R}$ is furthermore bounded. Although we have obtained interesting characterisations of the emptiness of $\mathcal{R} \cap \mathcal{Y}$, it is hard to get more theoretical knowledge from those coarse primal and dual functionals.

In this thesis, we tackle the intersection problem through evaluations of the separating functional J, with two potential applications:

- if one finds $p \in X$ such that $J(p) < 0$, then (0.4.7) entails $\mathcal{R} \cap \mathcal{Y} = \emptyset$
- if one deduces enough information from multiple evaluations to prove that $J \geq 0$ everywhere, then (0.4.7) proves $\mathcal{R} \cap \mathcal{Y} \neq \emptyset$.

Notice that while the first option requires only local information on J, the second demands global conditions, which is much more difficult to obtain. In both cases, a rigorous evaluation of J is called for: since a closed formula for J is rarely available in control problems, we consider numerical approximations of J, at the additional cost of then certifying bounds on the error between J and its proxy J_d – note that J_d may not be defined on the whole of X.

When numerically computing J, two kinds of error arise: approximation and rounding errors. Firstly, *approximation errors*: in the control problems we shall consider, these will boil down to discretisation errors of (partial) differential equations in time (and space for PDEs). To properly evaluate J, one then needs to precisely bound the discretisation error, that is, provide an *explicit* error function e_d such that

$$\forall\,(p_f, p_{fh}), \qquad |J(p_f) - J_d(p_{fh})| \leq e_d(p_f, p_{fh}). \tag{0.4.10}$$

This means that, in addition to providing an order of convergence of J_d towards J as the discretisation parameters tend to 0, one needs to ensure that the constants have a known closed form, and are furthermore as optimised as possible.

The second kind of error that appears when computing numerical approximations is *rounding errors* $e_r(p_{fh})$ usually arising from finite bit representation. These can be tedious to bound, as they may depend on a wide range of parameters independent of $\mathcal{R}$ and $\mathcal{Y}$: the computer, coding language and infrastructure are only some of them. To remedy this and rigorously bound rounding errors, we use in this thesis the MATLAB library INTLAB developed by Professor S. Rump [Rum99]. This library encodes interval arithmetic: a rigorous numerics technique that propagates worst-case rounding errors using intervals – see Section I.4 for a detailed account of its methods.

Ultimately, the goal is to provide rigorous bounds $e_d(p_f, p_{fh})$ on discretisation errors

$$J(p_f) \in [J_d(p_{fh}) - e_d(p_f, p_{fh}), J_d(p_{fh}) + e_d(p_f, p_{fh})], \tag{0.4.11}$$

and bounds $e_r(p_{fh})$ on rounding errors

$$J_d(p_{fh}) \in [\tilde{J}_d(p_{fh}) - e_r(p_{fh}),\ \tilde{J}_d(p_{fh}) + e_r(p_{fh})], \tag{0.4.12}$$

where $\tilde{J}_d(p_{fh})$ is the numerical evaluation of $J_d(p_{fh})$ subject to rounding errors, to finally prove the inclusion

$$J(p_f) \in [\tilde{J}_d(p_{fh}) - e_d(p_f, p_{fh}) - e_r(p_{fh}),\ \tilde{J}_d(p_{fh}) + e_d(p_f, p_{fh}) + e_r(p_{fh})]. \tag{0.4.13}$$

From there, the objective is then to extract whatever information it entails regarding the emptiness of $\mathcal{R} \cap \mathcal{Y}$.

0.4.3 Application to reachability analysis

Let us now apply those results to reachability analysis. Recall the control problem

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{for a.e. } t \in [0,T], \\ y(0) = y_0 \in \mathcal{Y}_0, & \\ u(t) \in \mathcal{U} & \text{for a.e. } t \in [0,T], \end{cases} \tag{$\mathcal{S}$}$$

and its solution

$$y : t \mapsto S_t y_0 + L_t u. \tag{0.4.14}$$

A target set $\mathcal{Y}_f \subset X$ is thus $\mathcal{U}$-reachable from $\mathcal{Y}_0$ in time T if and only if

$$\exists\, u \in E_{\mathcal{U}}, \exists\, (y_0, y_f) \in \mathcal{Y}_0 \times \mathcal{Y}_f, \quad S_T y_0 + L_T u = y_f, \tag{0.4.15}$$

or equivalently if

$$L_T E_{\mathcal{U}} \cap (\mathcal{Y}_f - S_T \mathcal{Y}_0) \neq \emptyset, \tag{0.4.16}$$

where

$$\mathcal{Y}_f - S_T \mathcal{Y}_0 = \{y_f - S_T y_0,\ (y_0, y_f) \in \mathcal{Y}_0 \times \mathcal{Y}_f\}. \tag{0.4.17}$$

Under the assumptions that

$$\begin{cases} L_T E_{\mathcal{U}} & \text{is nonempty, convex, closed and bounded} \\ \mathcal{Y}_f - S_T \mathcal{Y}_0 & \text{is nonempty, convex and closed,} \end{cases} \tag{0.4.18}$$

we can thus apply all results from the previous sections to study the reachability of $\mathcal{Y}_f$. Notice more tractable sufficient conditions are the following ones:

$$\begin{cases} \mathcal{U} \subset U & \text{is nonempty, closed, bounded and convex,} \\ \mathcal{Y}_f \subset X & \text{is nonempty, closed and convex,} \\ \mathcal{Y}_0 \subset X & \text{is nonempty, closed, bounded and convex.} \end{cases} \tag{0.4.19}$$

For the control constraints, one might even consider the sufficient assumption that $E_{\mathcal{U}}$ is nonempty and weakly compact. Indeed, it is standard that $L_T E_{\mathcal{U}}$ is weakly compact, and its convexity follows from profound results – see [LM86, Theorem 1A, Theorem 3 and Lemma 4A in Section 2.2] for a proof for finite dimensional systems, or [LY12, Chapter 7, Section 5.1] in infinite-dimension.

As for $\mathcal{Y}_f - S_T \mathcal{Y}_0$, it only needs convexity and closedness to ensure separation with $L_T E_{\mathcal{U}}$, which requires boundedness of $\mathcal{Y}_0$ as well to ensure the closedness of $S_T \mathcal{Y}_0$.

The separating functional in the control case writes

$$J : \begin{cases} X & \to \mathbb{R} \\ p_f & \mapsto \sigma_{L_T E_{\mathcal{U}}}(p_f) + \sigma_{\mathcal{Y}_f - S_T \mathcal{Y}_0}(-p_f), \end{cases} \tag{0.4.20}$$

and we deduce from the Subsection 0.4.1 the following theorem:

Theorem 0.3. Considering the linear control system ($\mathcal{S}$), where $\mathcal{U}$ denote nonempty, convex and weakly compact constraints, and $\mathcal{Y}_f - S_T \mathcal{Y}_0$ is nonempty, closed and convex, it holds that

$$\mathcal{Y}_f \text{ is } \mathcal{U}\text{-reachable from } \mathcal{Y}_0 \text{ in time } T \quad \Longleftrightarrow \quad \forall\, p_f \in X,\ J(p_f) \geq 0. \tag{0.4.21}$$

0.4.4 Computation of J

In general control problems, since the reachable set is not known, neither is the support function $\sigma_{L_T E_\mathcal{U}}$. Fortunately, a well-known property of support functions, along with the linearity of L_T, reduces the knowledge of the support function of $L_T E_\mathcal{U}$ to that of $E_\mathcal{U}$ – which is none other but the support function of $\mathcal{U}$ integrated in time:

$$\forall\, p_f \in X, \quad \sigma_{L_T E_\mathcal{U}}(p_f) = \sigma_{E_\mathcal{U}}(L_T^* p_f) = \int_0^T \sigma_\mathcal{U}(L_T^* p_f(t))\,\mathrm{d}t. \tag{0.4.22}$$

Other properties allowing us to split $\sigma_{\mathcal{Y}_f - S_T \mathcal{Y}_0}$ in two, we finally obtain the convenient expression

$$\forall\, p_f \in X, \quad J(p_f) = \int_0^T \sigma_\mathcal{U}(L_T^* p_f(t))\,\mathrm{d}t + \sigma_{\mathcal{Y}_0}(S_T^* p_f) + \sigma_{\mathcal{Y}_f}(-p_f). \tag{0.4.23}$$

Provided that the support functions of the constraints, initial and target sets are known (which is common for most standard cases) this expression is approximable, which paves the way to rigorous numerical computation as described in Subsection 0.4.2. Let us now introduce a general discretisation method for J: notice that considering the adjoint equation

$$\begin{cases} p'(t) + A^* p(t) = 0 & \text{for a.e. } \ t \in [0,T] \\ p(T) = p_f, \end{cases} \tag{0.4.24}$$

J can once again be rewritten as

$$\forall\, p_f \in X, \quad J(p_f) = \int_0^T \sigma_\mathcal{U}(B^* p(t))\,\mathrm{d}t + \sigma_{\mathcal{Y}_0}(p(0)) + \sigma_{\mathcal{Y}_f}(-p_f). \tag{0.4.25}$$

This reduces the discretisation of J to the discretisation of both a time integral and a linear differential equation: since for standard control constraints (bounded and non-singleton) $\sigma_\mathcal{U}$ has only Lipschitz regularity, we only consider first order methods to compute the time integral – and thus, if a discretisation of 0.4.24 is needed, a first-order discretisation in time is sufficient as well. The actual discretisation depending on the context – finite- or infinite-dimensional – we shall detail it in the following sections.

The following subsections summarise the main contributions of Chapters II, III and IV of this thesis respectively.

0.4.5 Computer-assisted proofs of non-reachability for finite-dimensional control systems

Non-reachability of finite-dimensional control systems is treated in Chapter II, which is a nearly identical transcript of the paper *Computer-assisted proofs of non-reachability for finite-dimensional linear control systems* co-written with Camille Pouchol, Yannick Privat and Christophe Zhang [Has+24], published in *SIAM Journal of Control and Optimisation* in September 2025.

In this article, we consider finite-dimensional control problems

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{for a.e. } \ t \in [0,T], \\ y(0) = y_0 \in \mathbb{R}^n, & \\ u(t) \in \mathcal{U} & \text{for a.e. } \ t \in [0,T], \end{cases} \tag{$\mathcal{S}$}$$

where $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and $B \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ are matrices and $\mathcal{U} \subset \mathbb{R}^m$ is compact (with bound $M > 0$). In particular, semigroups $(S_t)_{t\geq 0}$ reduce to matrix exponentials $(e^{tA})_{t\geq 0}$ and space discretisation is not needed.

This article exploits the contrapositive of Theorem 0.3, that is, for any target set $\mathcal{Y}_f$ nonempty, closed and convex:

$$\mathcal{Y}_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T \quad \Longleftrightarrow \quad \exists\, p_f \in \mathbb{R}^n,\ J(p_f) < 0. \tag{0.4.26}$$

Such a p_f will then be referred to as a *dual certificate of non-reachability*. In contrast to previous methods considered in the literature, the purpose is not to produce approximations of the reachable set, but rather to provide rigorous guarantees that the given target set is not reached. This simpler goal allows us to develop a general methodology adapted to a wide range of finite-dimensional linear control problems. This methodology relies on three consecutive steps:

1. the discretisation of J and explicit bound on the discretisation error e_d
2. the minimisation of $J_{\Delta t}$ using primal-dual algorithms, until a dual certificate p_f is found such that $J_{\Delta t}(p_f) < 0$
3. the rigorous computation of $J_{\Delta t}(p_f)$ with interval arithmetic to bound rounding errors e_r, and the conclusion that $J(p_f) \leq J_{\Delta t}(p_f) + e_r(p_f) + e_d(p_f) < 0$.

Another essential by-product of this method is the proof of lower bounds on minimal times of reachability; this is a consequence of the following basic result:

> **Lemma 0.4** (II.7)**.** Assume that $\mathcal{U} \cap \ker(B) \neq \emptyset$, and suppose either $y_0 = 0$ or $y_f = 0$. If y_f is not $\mathcal{U}$-reachable from y_0 in time T, then it is not reachable for any $\tilde{T} \leq T$ either. Consequently, denoting
>
> $$T^\star(y_0, y_f, \mathcal{U}) = \inf \{T > 0,\ y_f \text{ is } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T\} \in [0 + \infty],$$
>
> we have $T^\star(y_0, y_f, \mathcal{U}) \geq T$.

For the discretisation of J, we consider several cases, depending on the matrix A. Firstly, if a closed formula exists for the solution of (0.4.24) – for instance using a Jordan-Chevalley decomposition – the discretisation of J boils down to a simple discretisation of the integral. This discretisation is made using a simple rectangle rule – recall that the support function does not usually have more that Lipschitz regularity. Proposition II.9 and its Corollary II.10 study discretisations of the type

$$J_{\Delta t}(p_f) = \sum_{k=0}^{N_0-1} \sigma_{\mathcal{U}} \left(B^* e^{k\Delta t A^*} p_f\right) + \left\langle y_0, e^{TA^*} p_f \right\rangle + \sigma_{\mathcal{Y}_f}(-p_f), \tag{0.4.27}$$

and prove estimates of the form

$$|J(p_f) - J_{\Delta t}(p_f)| \leq \frac{1}{2} \Delta t M T \varphi(T) \|B\| \|A^* p_f\|, \tag{0.4.28}$$

where φ is typically the product of a polynomial and an exponential in T, depending on the eigenvalues of A.

When less information is known about A, a discretisation of the adjoint equation (0.4.24) is necessary. Under the assumption that A is negative semidefinite, we consider an implicit Euler scheme:

$$\begin{cases} p_{N_0} = p_f \\ (\mathrm{Id} - \Delta t A^*) p_k = p_{k+1}, \quad \forall\, k \in \{0, \ldots, N_0 - 1\}. \end{cases} \tag{0.4.29}$$

Proposition II.14 then yields the following estimate for the full discretisation of the separating functional:

$$|J(p_f) - J_{\Delta t}(p_f)| \leq \Delta t \|A^* p_f\| \left(TM\|B\| + \frac{1}{2}\|y_0\|\right). \tag{0.4.30}$$

To use Theorem 0.3, one first has to find an appropriate dual certificate of non-reachability p_f, which is done by “minimising” $J_{\Delta t}$. To do so, we made use of the natural duality of the problem, as displayed in (0.4.9), to employ a primal-dual algorithm [CP11]. The method has then successfully been applied to three examples in Section II.4:

- in Subsection II.4.2, a 2-dimensional toy example modelling the movement of a streetcar. This example's reachable set has a closed expression, which allows us to showcase the effectiveness of our approach
- in Subsection II.4.3, a more involved 4-dimensional model of a spacecraft rendezvous with complex control bounds and multiple target shapes. See for example Figure II.4 that demonstrates the ability of the method to compute accurate lower bounds of minimal reachability times
- in Subsection II.4.4, a more abstract example studying the scalability of the method in higher dimensions.

Each numerically assisted result is made using fully rigorous interval arithmetic encoding, and the dual certificates p_f and enclosing values of $J(p_f)$ are provided.

0.4.6 Computer-assisted proofs of non-reachability for parabolic control systems

Chapter III presents the results of an article co-written with Camille Pouchol, Yannick Privat and Christophe Zhang, which is currently being finalised for submission to a peer-reviewed journal.

In this chapter, we extend the method to study the non-reachability of linear parabolic control problems of the form

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{for a.e. } \ t \in [0,T], \\ y(0) = y_0 \in X, \\ u(t) \in \mathcal{U} & \text{for a.e. } \ t \in [0,T], \end{cases} \tag{S}$$

where $V \subset X \subset V'$ a Gelfand triple, with V and X infinite dimensional Hilbert spaces. Let $A : V \to V'$ be an operator, with its domain defined as $\mathcal{D}(A) = \{x \in V, Ax \in X\}$. Let U be another Hilbert space, with $B : U \to X$ a bounded operator. We assume that $\mathcal{U} \subset U$ is a nonempty, convex, closed and bounded (with bound $M > 0$), and consider nonempty, closed and convex target sets $\mathcal{Y}_f$.

Similarly to the finite-dimensional case, we aim at using (0.4.26) to create a computer-assisted methodology to prove the non-reachability of target sets and compute certified lower-bounds of minimal times of reachability for a wide array of parabolic control problems. However, to our knowledge, no computer-assisted proofs pertaining to the reachability analysis of PDE controlled systems have been considered in the literature.

The infinite-dimensionality of the state space significantly complicates the study of the separating functional J, for it requires its space-discretisation in addition to its time-discretisation. These call for many theoretical results from operator and semigroup theories as well as approximation theory: see Sections I.1 and I.3 for a brief introduction of those theories.

In a first part (Section III.2), we describe a space- and time-discretisation based on a variational formulation of the adjoint equation (0.4.24). We assume continuity and coercivity of operator $-A^*$, and consider a family of finite-dimensional subspaces $V_h \subset V$ indexed by a discretisation parameter $h > 0$. We also assume that V_h satisfies the standard approximation assumption

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

with $C_0 \geq 0$ a constant that will need to be explicit for later purposes. Over V_h, we discretise A into A_h, and solve the adjoint equation (0.4.24) in time using a Euler implicit scheme. Using approximation theory and quantitative results for accretive operators, we obtain the following explicit discretisation error bound (see Proposition III.9):

$$\|p(t_n) - p_{h,n}\| \le C_1\|p_f - p_{fh}\| + (C_2h^2 + C_3\Delta t)\|A^*p_f\|, \tag{0.4.31}$$

where $\Delta t > 0$ and $h > 0$ are the time and space discretisation parameters, and the constants are explicit and only depend on the continuity and coercivity constants of $-A^*$ and on C_0. A fully discretised version of J is the following:

$$J_{\Delta t,h}(p_{fh}) = \sum_{k=0}^{N_0-1} \sigma_{\mathcal{U}}\left(B^*(\mathrm{Id} - \Delta t A_h)^{-k} p_{fh}\right) + \left\langle y_0, (\mathrm{Id} - \Delta t A_h)^{-N_0} p_{fh}\right\rangle + \sigma_{\mathcal{Y}_f}(p_{fh}). \tag{0.4.32}$$

An explicit error bound on the functional J is subsequently derived from this estimate (see Theorem III.11):

$$\begin{aligned}|J(p_f) - J_{\Delta t,h}(p_{fh})| \le \left(\tfrac{1}{2}M\|B\|T\Delta t + (\|y_0\| + M\|B\|T)(C_2h^2 + C_3\Delta t)\right)\|A^*p_f\| \\ + (C_4(\|y_0\| + M\|B\|T) + \|y_f\|)\|p_f - p_{fh}\|.\end{aligned} \tag{0.4.33}$$

Notice that in both estimates (0.4.31) and (0.4.33), we have $p_{fh} \in V_h$ and $p_f \in \mathcal{D}(A^*)$. Recall that we do not necessarily have $V_h \subset \mathcal{D}(A^*)$, and thus any $p_{fh} \in V_h$ obtained through a minimisation process must be interpolated into $p_f \in \mathcal{D}(A^*)$ in order to apply (0.4.33) and guarantee the value of $J(p_f)$. Therefore, in addition to the methodology described in Subsection 0.4.5, one has to consider the following steps:

- before minimising on V_h the discretised functional $J_{\Delta t,h}$ defined in (0.4.32), estimate ($\mathcal{V}_1$) must be proved for the discretisation space V_h.
- to compute the final discretisation errors, one must interpolate p_{fh} into $p_f \in \mathcal{D}(A^*)$ and check that $J(p_f) < 0$ using previous estimate (0.4.33) and interval arithmetic.

In Section III.3, we apply this general recipe to prove non-reachability results on two heat equation based examples: in both cases, we choose a space discretisation that uses $\mathbb{P}_1$ finite elements, and then interpolate the minimiser of J_d using cubic splines – see Section I.3 for details about these spaces.

Using those results, we first study the 1D heat equation with zero Dirichlet boundary conditions: a toy case with symmetric L^2 constraints and $B = \mathrm{Id}$ is considered, showcasing the precision of results for simple targets. We then consider more involved examples with restricted control domains ($B = \chi_\omega$) and nonnegative L^∞ control constraints, proving the non-reachability and computing lower-bounds of minimal reachability time for various targets – see for example Figures III.2 and III.3.

We then consider a coupled heat equation example with a control applied to only the second equation – control satisfying asymmetric L^∞ constraints. After the necessary proof of the approximation result ($\mathcal{V}_1$), we provide computer-assisted proofs of the non-reachability of a ball centred at 0 – see Figure III.4 for an illustration of the initial state and dual certificate. We also prove the non-reachability of an unbounded target set of the form $\{y_f\} \times L^2(0,1)$.

In all considered examples, the computer-assisted proofs use rigorously encoded interval arithmetic. The dual certificates and enclosing values of J are provided and discussed with regard to all parameters. The discretisation errors and rounding errors are also shown and when necessary, their respective importance is also discussed.

0.4.7 Computer-assisted proofs of reachability for finite-dimensional control systems

Chapter IV is devoted to computer-assisted proofs of reachability in finite-dimensional linear systems. It presents ongoing work aiming to:

- prove the reachability of a target set $\mathcal{Y}_f$

- provide over-approximations (which contain the reachable set) and under-approximations (included in it) of the reachable set.

In order to develop those, we rely on the computation of the separating functional J described above, as well as on the discretisations studied in Chapter II. Since each computation of J proves the non-reachability of a whole half-space, the computation of J at several points well distributed around the origin permits the inclusion of the reachable set inside a bounded intersection of half-spaces – a bounded polytope. Increasing the number of evaluations of J then allows for tighter over-approximations of the reachable set.

Computing J also permits to prove the reachability of some targets: for a given polytopal over-approximation P of the reachable set, recall that each facet of this polytope is contained in a supporting hyperplane of the reachable set. Each facet thus contains at least one point of the reachable set. One can then compute the intersection $\mathcal{I}(P)$ of all convex sets with a point in each facet of P, which is then necessarily included in the reachable set. If one has computed the value of J at sufficiently many points, then this intersection $\mathcal{I}(P)$ is nonempty (see Figure IV.2). Increasing the number of hyperplanes then increases the size of this inner-approximation. Furthermore, one can choose where to evaluate J in order to "guide" the expansion of $\mathcal{I}(P)$ towards a target point and therefore conclude upon its reachability status.

Another method to compute under-approximations of the reachable set requires additional information on the computation of J: recall that computation of $J(p_f)$ for a given p_f involves the computation of the support function of the control constraints. Usually, this computation is done using the knowledge of the control maximising the scalar product. One can then calculate the state reached using this control, and computing the convex hull of all such points allows for a more precise under-approximation of the reachable set (see Figure IV.4). Similarly, one can adjust iteratively the evaluation points of J to study the reachability of a specific target.

Most evidently, such approximations of the reachable set require an extensive use of rigorous numerics to compute them. Accounting for discretisation and rounding errors significantly reduces the size of the under-approximations (see for instance Figures IV.8 and IV.9). On another note, many algorithms need to be adapted to interval (or affine) arithmetic: for instance, one has to switch between half-spaces to vertex definitions of polytopes, or to compute under approximations of convex hulls. These are challenging problems which, to our knowledge, remain unsolved. Chapter IV does not aim to provide a comprehensive treatment of the subject; rather, it introduces the main problems and outlines ideas that may contribute to their resolution.

0.4.8 Outlook on future work and outline of the thesis

Non-reachability in finite dimension. Chapter II introduces the method for finite-dimensional linear control systems: while we consider a large panel of linear systems, many remain to be considered. For instance, we believe that the method could be easily extended to non-homogeneous non-autonomous systems $y'(t) = A(t)y(t) + B(t)u(t) + v(t)$, at the cost of higher discretisation errors. Similarly, time-dependent control constraints could be considered, as well as control constraints for which the support function is not known explicitly – a new approximation would then need to be quantified. Another lead would be to find regularisations of the separating functional, allowing for higher order discretisation schemes.

The next major step would be to extend the method to nonlinear finite-dimensional systems. As it is, it depends highly on the linearity of L_T and on the convexity of the reachable set, such an extension would thus require some major modifications. Until now we have only considered support hyperplanes, one could consider tailoring those separating surfaces to each nonlinear problem, for example using optimal control characterisations of adjoint systems such as the Pontryagin Maximum Principle. Another lead could be to lift the system using Koopman's operator and consider the linear transport PDE whose resolution is equivalent to the one of nonlinear ODE, and apply the method to the linear PDE in a manner similar to the one used in Chapter III.

Non-reachability in infinite-dimension. The question of linear PDEs is much more complex, for it requires space-discretisation in addition to time-discretisation. In particular, we have only considered continuous and coercive operators for parabolic equations: extensions to other operators would be a first step for a more versatile methodology. Similarly to finite-dimensional systems, the crux of the method is to obtain explicit and small discretisation errors: an extension to hyperbolic PDEs would thus require a careful analysis of numerical schemes to avoid discretisation errors blowing up with respect to the final time. As for finite-dimensional systems, a simpler goal would be to tackle different parabolic systems, for instance including time-dependent components or more complicated boundary conditions. Studying boundary controls would as well be an interesting problem, made complex by the need of explicit discretisation errors for which the unboundedness of operator B is problematic.

Reachability in finite-dimension. Another goal of future research is to develop a methodology for computer-assisted proofs of reachability in finite-dimension. This is more complex, for it involves computing over-approximations of the reachable set, which is computationally heavy, especially using interval arithmetic. A related task is to compute under-approximations of the reachable set, which is even more computationally expensive. Since those approximations only rely on the computation of the separating functional, these methods could be extended in the same way as the non-reachability study to time-dependent systems. However, both nonlinear systems and infinite-dimensional systems seem unattainable: the first because the reachable set is nonconvex, and the latter because it could only provide finite-dimensional under-approximations of the infinite-dimensional reachable set.

Outline of the thesis. This thesis is divided into four parts: Chapter I introduces many useful frameworks that will be needed for the main parts of the thesis: operator and semigroup theories, convex analysis, approximation theory and rigorous numerics. In Chapter II, we formalise the method described above for finite-dimensional control problems, and apply it to the study of non-reachability. In Chapter III, we extend this method to tackle the non-reachability of infinite-dimensional linear parabolic control problems. Finally, Chapter IV presents ongoing work to obtain computer-assisted proofs of reachability for finite-dimensional control problems.

0.5 Résumé de la thèse

In this section, we shall translate the contents of Section 0.4 into French.

Cette thèse est vouée à l'étude de systèmes linéaires autonomes de contrôle : nous considérons un opérateur A générateur d'un semi-groupe fortement continu, ainsi qu'un opérateur continu B, formant le système

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{pour presque tout } \ t \in [0,T], \\ y(0) = y_0 \in \mathcal{Y}_0, & \\ u(t) \in \mathcal{U} & \text{pour presque tout } \ t \in [0,T]. \end{cases} \tag{$\mathcal{S}$}$$

Ici, $\mathcal{U}$ représente des contraintes sur le contrôle, $\mathcal{Y}_0 \subset X$ un ensemble de conditions initiales et $\mathcal{Y}_f \subset X$ un ensemble cible. En général, nous faisons les hypothèses suivantes :

$$\begin{cases} \mathcal{U} \subset U & \text{est non vide, fermé, convexe et borné} \\ \mathcal{Y}_f \subset X & \text{est non vide, fermé, convexe} \\ \mathcal{Y}_0 \subset X & \text{est non vide, fermé, convexe et borné.} \end{cases} \tag{0.5.1}$$

Nous considérons également des hypothèses légèrement plus faibles, voir (0.5.18).

Dans cette thèse, un ensemble cible $\mathcal{Y}_f$ est dit $\mathcal{U}$-atteignable depuis $\mathcal{Y}_0$ en temps T s'il existe un couple condition initiale - cible $(y_0, y_f) \in \mathcal{Y}_0 \times \mathcal{Y}_f$ tel que y_f est $\mathcal{U}$-atteignable en temps T depuis y_0. Cette définition est non conventionnelle, mais coïncide avec la définition usuelle quand $\mathcal{Y}_0$ et $\mathcal{Y}_f$ sont des singletons.

L'objectif central de cette thèse est de répondre à la question suivante :

> Pour un système linéaire de contrôle soumis à des contraintes bornées $\mathcal{U}$ sur le contrôle, des ensembles initial $\mathcal{Y}_0$ et cible $\mathcal{Y}_f$, et un temps final T, $\mathcal{Y}_f$ est-il $\mathcal{U}$-atteignable en temps T depuis $\mathcal{Y}_0$?

Nous limiterons notre étude à des contrôles en boucle ouverte, et nous focaliserons principalement sur des preuves de non-atteignabilité, ce qui en particulier empêche l'atteignabilité en boucle fermée.

Le résumé de cette thèse est divisé comme suit :

- Les sous-sections 0.5.1 et 0.5.2 se concentrent sur un problème d'analyse convexe : nous étudions une fonctionnelle J caractérisant la séparation entre deux ensembles convexes, et relions son évaluation à l'existence d'un point commun aux deux ensembles
- Ensuite, les sous-sections 0.5.3 et 0.5.4 introduisent une méthodologie appliquant J à l'étude d'atteignabilité pour des systèmes de contrôle linéaires
- Les sous-sections 0.5.5, 0.5.6 et 0.5.7 présentent les principaux résultats de cette thèse, chaque sous-section étant dédiée à un chapitre de cette thèse (chapitres II, III et IV)
- Finalement, la sous-section 0.5.8 présente les perspectives de cette thèse ainsi que son plan.

0.5.1 Analyse convexe

Nous commençons donc par présenter la méthodologie d'un point de vue d'analyse convexe : voir la section I.2 pour une présentation plus détaillée de cette théorie. Soit X un espace de Hilbert et deux ensembles convexes non vides $\mathcal{R}, \mathcal{Y} \subset X$. Nous nous concentrons ici sur la question suivante :

$$\text{Est } \mathcal{R} \cap \mathcal{Y} \text{ vide ?} \tag{0.5.2}$$

C'est une question classique d'analyse convexe, qui peut être reformulée comme un problème d'optimisation convexe :

$$\pi := \inf_{x \in X} \delta_{\mathcal{R}}(x) + \delta_{\mathcal{Y}}(x), \tag{$\mathcal{P}$}$$

où δ_C est l'indicatrice convexe d'un ensemble C, valant 0 sur C et $+\infty$ en dehors. Ce problème d'optimisation quelque peu brut permet de répondre à la question (0.5.2) dans le sens suivant :

$$\mathcal{R} \cap \mathcal{Y} \neq \emptyset \quad \Longleftrightarrow \quad \pi = 0. \tag{0.5.3}$$

Il va de soi que ce problème d'optimisation pourrait être régularisé de multiples façons tout en conservant cette propriété, mais sa forme simple actuelle nous suffira ici.

Une autre manière d'aborder (0.5.2) est d'utiliser un argument de séparation tel que le théorème d'Hahn-Banach : en supposant que

$$\begin{cases} \mathcal{R} \text{ et } \mathcal{Y} \text{ sont non vides, fermés et convexes} \\ \mathcal{R} \text{ ou } \mathcal{Y} \text{ est borné,} \end{cases} \tag{0.5.4}$$

on a l'équivalence

$$\mathcal{R} \cap \mathcal{Y} = \emptyset \quad \Longleftrightarrow \quad \exists p \in X, \sup_{x \in \mathcal{R}} \langle p, x \rangle < \inf_{y \in \mathcal{Y}} \langle p, y \rangle. \tag{0.5.5}$$

Posons $\sigma_C : p \mapsto \sup_{x \in C} \langle p, x \rangle$ la fonction de support d'un ensemble convexe fermé C, et

$$J : \begin{cases} X & \to \mathbb{R} \\ p & \mapsto \sigma_{\mathcal{R}}(p) + \sigma_{\mathcal{Y}}(-p). \end{cases} \tag{0.5.6}$$

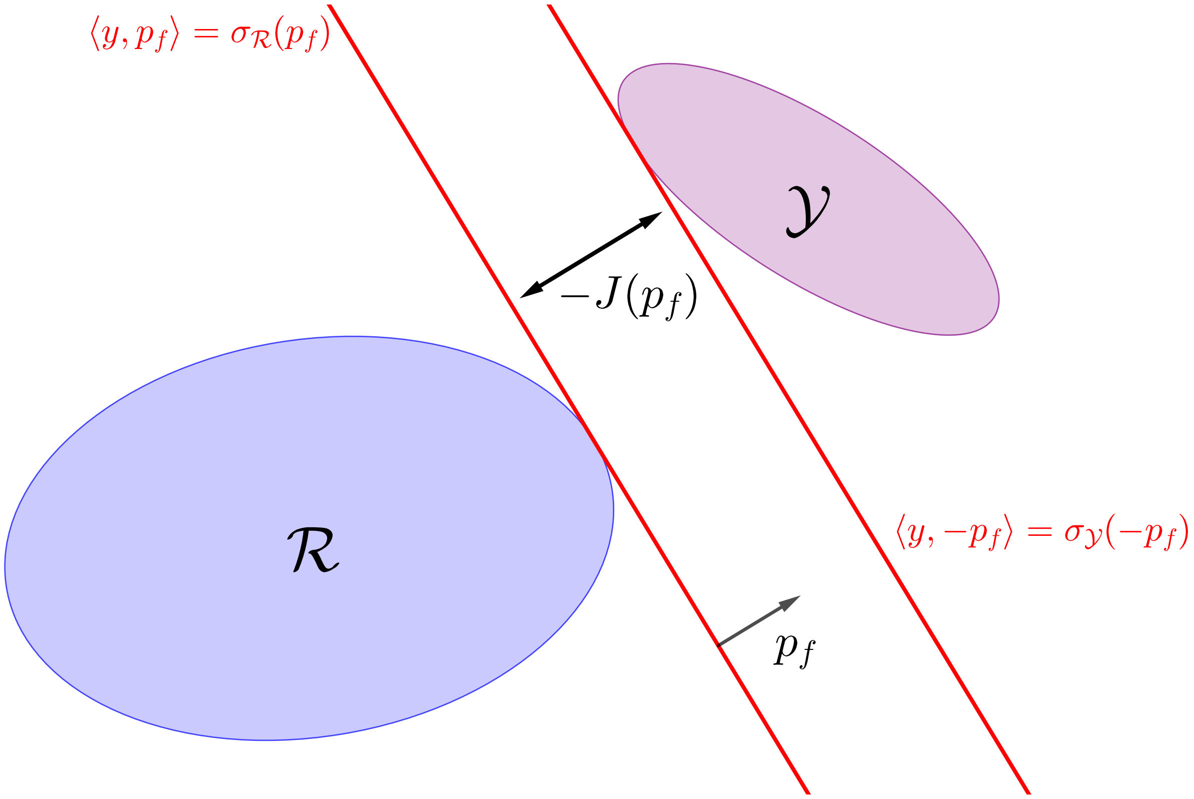

$\langle y, p_f \rangle = \sigma_{\mathcal{R}}(p_f)$
$\mathcal{Y}$
$-J(p_f)$
$\mathcal{R}$
$\langle y, -p_f \rangle = \sigma_{\mathcal{Y}}(-p_f)$
p_f

FIGURE 0.3 : Séparation de $\mathcal{R}$ et $\mathcal{Y}$ avec J.

J quantifie la séparation entre $\mathcal{R}$ et $\mathcal{Y}$ (voir l'illustration en Figure 0.3), elle sera parfois appelée *fonctionnelle de séparation*. Comme J est 1-homogène, le problème de minimisation

$$d := \inf_{p \in X} J(p) \tag{$\mathcal{D}$}$$

peut uniquement avoir deux valeurs : 0 ou $-\infty$. Sous les hypothèses (0.5.4), l'équivalence (0.5.3) tient également :

$$\mathcal{R} \cap \mathcal{Y} \neq \emptyset \quad \Longleftrightarrow \quad d = 0. \tag{0.5.7}$$

De fait, les deux problèmes de minimisation – le *problème primal* ($\mathcal{P}$) et le *problème dual* ($\mathcal{D}$) – sont liés par la dualité de Fenchel-Rockafellar : celle-ci garantit leur *dualité faible*, c'est-à-dire

$$\pi \geq -d. \tag{0.5.8}$$

Sous les hypothèses (0.5.4), la *dualité forte* découle du théorème de Fenchel-Rockafellar [Roc67] (voir le théorème I.24).

$$\pi = -d. \tag{0.5.9}$$

Au cours de cette thèse, nous mettrons à profit cette dualité, et tout particulièrement la fonctionnelle J, pour développer des preuves assistées par ordinateur pour certifier des approximations de J, puis pour en déduire des certificats de (non-)atteignabilité. Dans la prochaine sous-section, nous donnons quelques détails quant à l'évaluation rigoureuse de la fonctionnelle de séparation.

0.5.2 Calcul certifié de la fonctionnelle séparatrice

Par la suite, nous nous placerons sous les hypothèses (0.5.4), à savoir que $\mathcal{R}$ et $\mathcal{Y}$ sont des ensembles convexes, fermés et non vides, et que de plus $\mathcal{R}$ est borné. Malgré les équivalences entre l'existence d'un point dans $\mathcal{R} \cap \mathcal{Y}$ et les précédents problèmes de minimisation, l'absence de régularité de ceux-ci complique l'obtention d'informations supplémentaires sur les fonctionnelles primales et duales. Dans cette thèse, nous abordons cette question en évaluant la fonctionnelle J, avec deux objectifs :

- si l'on trouve $p \in X$ satisfaisant $J(p) < 0$, alors (0.5.7) implique que $\mathcal{R} \cap \mathcal{Y} = \emptyset$
- si au contraire suffisamment d'informations sont déduites de multiples évaluations de J pour prouver sa positivité sur X, alors (0.5.7) implique que $\mathcal{R} \cap \mathcal{Y} \neq \emptyset$.

Remarquons que si la première option ne nécessite qu'une information locale sur J, la seconde requiert une information globale, significativement plus complexe à obtenir à partir d'évaluations locales. Puisque dans les applications à la théorie du contrôle il n'existe généralement pas de formule explicite de J, il faut alors estimer rigoureusement J à l'aide d'approximations numériques. Cela nécessite notamment de borner explicitement les erreurs entre J et son approximation J_d ; ce dernier est potentiellement non défini sur X tout entier.

L'approximation numérique de J fait apparaître deux types d'erreurs distincts : des erreurs d'approximation d'une part, et d'arrondi d'autre part. Tout d'abord, les *erreurs d'approximation* consistent avant tout en des erreurs de discrétisation (en temps et/ou espace) d'équations aux dérivées ordinaires ou partielles. Cela revient donc à trouver une fonction d'erreur *explicite* e_d telle que

$$\forall\,(p_f, p_{fh}), \qquad |J(p_f) - J_d(p_{fh})| \leq e_d(p_f, p_{fh}). \tag{0.5.10}$$

En particulier, il faut donc non seulement calculer un ordre de convergence de J_d vers J avec les paramètres de discrétisation, mais également s'assurer que l'ensemble des constantes impliquées dans cette convergence ont une forme explicite et soient aussi optimisées que possible.

Le second type d'erreurs intervenant apparaît lors du calcul numérique de J_d : celui-ci requiert généralement des représentations en nombres flottants de précision finie, sujettes aux *erreurs d'arrondi*. Celles-ci dépendent d'un nombre important de paramètres indépendants de $\mathcal{R}$ et $\mathcal{Y}$ (ordinateur, langage et environnement de programmation par exemple) et sont donc complexes à borner. Dans cette thèse, nous bornons ces erreurs à l'aide de la bibliothèque de calcul rigoureux INTLAB codée sous MATLAB par le Professeur S. Rump [Rum99]. Cette bibliothèque met à disposition des codes d'arithmétique d'intervalles, une méthode permettant la propagation certifiée d'erreurs d'arrondi à l'aide d'intervalles – voir la section I.4 pour plus de détails sur ces méthodes.

L'objectif final est donc de fournir des bornes rigoureuses et explicites $e_d(p_f, p_{fh})$ des erreurs de discrétisation

$$J(p_f) \in [J_d(p_{fh}) - e_d(p_f, p_{fh}), J_d(p_{fh}) + e_d(p_f, p_{fh})], \tag{0.5.11}$$

et d'arrondi $e_a(p_{fh})$

$$J_d(p_{fh}) \in [\tilde{J}_d(p_{fh}) - e_a(p_{fh}),\ \tilde{J}_d(p_{fh}) + e_a(p_{fh})], \tag{0.5.12}$$

où $\tilde{J}_d(p_{fh})$ est l'évaluation numérique de $J_d(p_{fh})$ sujette aux erreurs d'arrondis. Il en découle que

$$J(p_f) \in [\tilde{J}_d(p_{fh}) - e_d(p_f, p_{fh}) - e_a(p_{fh}),\ \tilde{J}_d(p_{fh}) + e_d(p_f, p_{fh}) + e_a(p_{fh})], \tag{0.5.13}$$

dont il faut ensuite déduire des informations sur l'existence d'un élément dans $\mathcal{R} \cap \mathcal{Y}$.

0.5.3 Application à l'étude d'atteignabilité

Appliquons à présent cette méthode à l'analyse d'atteignabilité. Rappelons le système contrôlé

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{pour presque tout } t \in [0, T], \\ y(0) = y_0 \in \mathcal{Y}_0, & \\ u(t) \in \mathcal{U} & \text{pour presque tout } t \in [0, T], \end{cases} \tag{$\mathcal{S}$}$$

ainsi que sa solution

$$y : t \mapsto S_t y_0 + L_t u. \tag{0.5.14}$$

Il s'ensuit qu'un ensemble cible $\mathcal{Y}_f \subset X$ est $\mathcal{U}$-atteignable si et seulement si

$$\exists\, u \in E_{\mathcal{U}}, \exists\, (y_0, y_f) \in \mathcal{Y}_0 \times \mathcal{Y}_f, \quad S_T y_0 + L_T u = y_f, \tag{0.5.15}$$

ou de manière équivalente si

$$L_T E_{\mathcal{U}} \cap (\mathcal{Y}_f - S_T \mathcal{Y}_0) \neq \emptyset, \tag{0.5.16}$$

où

$$\mathcal{Y}_f - S_T \mathcal{Y}_0 = \{y_f - S_T y_0,\ (y_0, y_f) \in \mathcal{Y}_0 \times \mathcal{Y}_f\}. \tag{0.5.17}$$

Sous les hypothèses que

$$\begin{cases} L_T E_{\mathcal{U}} & \text{est non vide, convexe, fermé et borné} \\ \mathcal{Y}_f - S_T \mathcal{Y}_0 & \text{est non vide, convexe et fermé,} \end{cases} \tag{0.5.18}$$

nous pouvons appliquer les résultats des sous-sections précédentes pour étudier l'atteignabilité de $\mathcal{Y}_f$. Remarquons que des hypothèses suffisantes plus aisément vérifiables sont les suivantes :

$$\begin{cases} \mathcal{U} \subset U & \text{est non vide, convexe, fermé et borné,} \\ \mathcal{Y}_f \subset X & \text{est non vide, convexe, et fermé,} \\ \mathcal{Y}_0 \subset X & \text{est non vide, convexe, fermé et borné.} \end{cases} \tag{0.5.19}$$

De fait, une hypothèse suffisante plus faible sur les contraintes que $E_{\mathcal{U}}$ pourrait être considérée : il suffit que $E_{\mathcal{U}}$ soit non vide et faiblement compact. En effet, un résultat classique donne la faible compacité de $L_T E_{\mathcal{U}}$, et un autre plus profond garantit sa convexité – voir [LM86, Théorème 1A, théorème 3 et lemme 4A dans la section 2.2] pour une preuve en dimension finie, ou [LY12, Chapitre 7, section 5.1] en dimension infinie.

D'autre part, $\mathcal{Y}_f - S_T \mathcal{Y}_0$ ne doit être que convexe et fermé pour garantir la séparation avec $L_T E_{\mathcal{U}}$, mais cette dernière hypothèse requiert le caractère borné de $\mathcal{Y}_0$ pour s'assurer de la fermeture de $S_T \mathcal{Y}_0$.

Dans le contexte d'analyse d'atteignabilité, la fonctionnelle de séparation est donc

$$J : \begin{cases} X & \to \mathbb{R} \\ p_f & \mapsto \sigma_{L_T E_{\mathcal{U}}}(p_f) + \sigma_{\mathcal{Y}_f - S_T \mathcal{Y}_0}(-p_f), \end{cases} \tag{0.5.20}$$

et il s'ensuit de la sous-section 0.5.1 le théorème suivant :

> **Théorème 0.5.** Pour le système ($\mathcal{S}$), où $\mathcal{U}$ représente des contraintes non vides, convexes et faiblement compactes, et $\mathcal{Y}_f - S_T \mathcal{Y}_0$ est non vide, convexe et fermé, ce qui prouve que
>
> $$\mathcal{Y}_f \text{ est } \mathcal{U}\text{-atteignable depuis } \mathcal{Y}_0 \text{ en temps } T \quad \Longleftrightarrow \quad \forall\, p_f \in X,\ J(p_f) \geq 0. \tag{0.5.21}$$

0.5.4 Calcul de J

Pour des systèmes contrôlés, l'ensemble est généralement inconnu, et par conséquent il en va de même pour sa fonction support $\sigma_{L_T E_{\mathcal{U}}}$. Heureusement, une propriété bien connue des fonctions supports, couplée à la linéarité de L_T, nous permets de nous ramener à la fonction support des contraintes. Celles-ci étant indépendantes du temps, s'ensuivent les égalités suivantes :

$$\forall\, p_f \in X, \quad \sigma_{L_T E_{\mathcal{U}}}(p_f) = \sigma_{E_{\mathcal{U}}}(L_T^* p_f) = \int_0^T \sigma_{\mathcal{U}}(L_T^* p_f(t))\, \mathrm{d}t. \tag{0.5.22}$$

D'autres propriétés classiques nous permettent de séparer $\sigma_{\mathcal{Y}_f - S_T \mathcal{Y}_0}$ en deux pour obtenir finalement

$$\forall\, p_f \in X, \quad J(p_f) = \int_0^T \sigma_{\mathcal{U}}(L_T^* p_f(t))\, \mathrm{d}t + \sigma_{\mathcal{Y}_0}(S_T^* p_f) + \sigma_{\mathcal{Y}_f}(-p_f). \tag{0.5.23}$$

Sous l'hypothèse que les fonctions supports de $\mathcal{U}$, $\mathcal{Y}_f$ et $\mathcal{Y}_0$ sont connues (ce qui est le cas pour la plupart des cas d'études), cette expression est approximable, permettant alors une approche certifiée telle que décrite dans la sous-section 0.5.2. Nous présentons maintenant une méthode générale de discrétisation de J : remarquons qu'en introduisant l'équation adjointe

$$\begin{cases} p'(t) + A^* p(t) = 0 & \text{pour presque tout } \ t \in [0, T] \\ p(T) = p_f, \end{cases} \tag{0.5.24}$$

J se reformule en

$$\forall\, p_f \in X, \quad J(p_f) = \int_0^T \sigma_{\mathcal{U}}(B^* p(t))\, \mathrm{d}t + \sigma_{\mathcal{Y}_0}(p(0)) + \sigma_{\mathcal{Y}_f}(-p_f). \tag{0.5.25}$$

Cette formulation permet de réduire la discrétisation de J à celles d'une intégrale en temps et d'une équation différentielle linéaire. Celle-ci dépendant beaucoup du contexte (EDO ou EDP), nous la détaillerons dans les prochaines sous-sections. Néanmoins, remarquons que dans les cas standards où $\mathcal{U}$ n'est pas réduit à un singleton, $\sigma_{\mathcal{U}}$ est uniquement Lipschitz, ce qui nous conduira à ne considérer des discrétisations en temps d'ordre 1 uniquement.

Les sous-sections suivantes résument respectivement les contributions principales des chapitres II, III et IV.

0.5.5 Preuves assistées par ordinateur de non-atteignabilité pour des systèmes de contrôle de dimension finie

Dans le chapitre II, nous nous intéressons à la non-atteignabilité en dimension finie. Ce chapitre est une copie quasiment à l'identique de l'article *Computer-assisted proofs of non-reachability for finite-dimensional linear control systems* co-écrit avec Camille Pouchol, Yannick Privat et Christophe Zhang [Has+24], et accepté pour publication par *SIAM Journal of Control and Optimisation* en mai 2025.

Dans cet article, nous considérons le système contrôlé de dimension finie suivant :

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{pour presque tout } \ t \in [0, T], \\ y(0) = y_0 \in \mathbb{R}^n, & \\ u(t) \in \mathcal{U} & \text{pour presque tout } \ t \in [0, T], \end{cases} \tag{$\mathcal{S}$}$$

où $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ et $B \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ sont des matrices et $\mathcal{U} \subset \mathbb{R}^m$ est compact (borné par $M > 0$). En particulier, les semigroupes $(S_t)_{t \geq 0}$ sont donc des exponentielles de matrices $(e^{tA})_{t \geq 0}$ et aucune discrétisation en espace n'est nécessaire.

Cet article exploite la contraposée du théorème 0.3 qui s'écrit, pour $\mathcal{Y}_f$ non vide, fermée et convexe :

$$\mathcal{Y}_f \text{ n'est pas } \mathcal{U}\text{-atteignable depuis } y_0 \text{ en temps } T \iff \exists\, p_f \in \mathbb{R}^n, \ J(p_f) < 0. \tag{0.5.26}$$

Un tel p_f est alors appelé *certificat dual de non-atteignabilité*. Par rapport aux méthodes déjà considérées dans la littérature, l'objectif n'est pas de produire des approximations de l'ensemble atteignable mais de fournir des certifications rigoureuses que la cible n'est pas atteignable. Cet objectif plus simple nous permet de développer une méthodologie générale applicable à une multitude de problèmes de contrôle de dimension finie. Cette méthodologie repose sur trois étapes consécutives :

- la discrétisation de J et la majoration explicite des erreurs de discrétisation e_d

- la minimisation de J_{Δ_t} à l'aide d'algorithmes primal-dual, jusqu'à l'obtention d'un certificat dual p_f satisfaisant $J_{\Delta t}(p_f) < 0$
- le calcul certifié de $J_{\Delta t}(p_f)$ en utilisant de l'arithmétique d'intervalles pour majorer les erreurs d'arrondis e_a, pour arriver à la conclusion que $J(p_f) \leq J_{\Delta t}(p_f) + e_a(p_f) + e_d(p_f)$.

Une autre conséquence intéressante de cette méthode est la preuve de minorants de temps minimaux d'atteignabilité. Cela découle immédiatement du lemme classique suivant :

Lemme 0.6. Supposons que $\mathcal{U} \cap \ker(B) \neq \emptyset$, et que $y_0 = 0$ ou $y_f = 0$. Si y_f n'est pas $\mathcal{U}$-atteignable en temps T, alors il ne l'est pas non plus pour tout $\tilde{T} \leq T$. Par conséquent, en notant le temps minimal d'atteignabilité

$$T^\star(y_0, y_f, \mathcal{U}) = \inf \{T > 0,\ y_f \text{ est } \mathcal{U}\text{-atteignable depuis } y_0 \text{ en temps } T\} \in [0 + \infty],$$

celui-ci satisfait $T^\star(y_0, y_f, \mathcal{U}) \geq T$.

Pour discrétiser J, plusieurs cas sont distingués en fonction de la matrice A. Tout d'abord, si une formule explicite de la solution de (0.5.24) existe (à l'aide par exemple d'une décomposition de Dunford de A), la discrétisation de J se réduit à la discrétisation de l'intégrale en temps. Celle-ci est effectuée à l'aide d'une méthode des rectangles (rappelons que l'intégrande a uniquement une régularité Lipschitz). La proposition II.9 et son corollaire II.10 présentent des discrétisations du type

$$J_{\Delta t}(p_f) = \sum_{k=0}^{N_0-1} \sigma_{\mathcal{U}}\left(B^* e^{k\Delta t A^*} p_f\right) + \left\langle y_0, e^{TA^*} p_f \right\rangle + \sigma_{\mathcal{Y}_f}(-p_f), \tag{0.5.27}$$

montrant des résultats de la forme de

$$|J(p_f) - J_{\Delta t}(p_f)| \leq \frac{1}{2} \Delta t M T \varphi(T) \|B\| \|A^* p_f\|, \tag{0.5.28}$$

où φ est typiquement le produit d'un polynôme et d'une exponentielle en T, dépendant des valeurs propres de A.

Si la solution de (0.5.24) n'est pas connue, sa discrétisation est nécessaire. Sous l'hypothèse que A est semidéfinie négative, nous considérons un schéma d'Euler implicite (rétrograde en temps) :

$$\begin{cases} p_{N_0} = p_f \\ (\operatorname{Id} - \Delta t A^*) p_k = p_{k+1}, \quad \forall\, k \in \{0, \dots, N_0 - 1\}. \end{cases} \tag{0.5.29}$$

Pour cette discrétisation, la Proposition II.14 montre alors l'erreur de discrétisation suivante :

$$|J(p_f) - J_{\Delta t}(p_f)| \leq \Delta t \|A^* p_f\| \left(TM\|B\| + \frac{1}{2}\|y_0\| \right). \tag{0.5.30}$$

Pour mettre à profit le théorème 0.5, il faut ensuite trouver un certificat dual de non-atteignabilité p_f, ce que nous faisons en "minimisant" $J_{\Delta t}$. Pour cela, nous utilisons l'algorithme primal-dual décrit dans [CP11], qui utilise la dualité intrinsèque du problème décrite en (0.5.9) pour fournir une convergence rapide. Cette méthode est finalement appliquée avec succès sur trois exemples, dans la section II.4 :

- dans la sous-section II.4.2, nous considérons un exemple jouet en dimension 2 modélisant le mouvement d'un Tramway. L'ensemble atteignable de cet exemple étant explicite, nous pouvons l'utiliser pour démontrer l'efficacité de notre approche.
- dans la sous-section II.4.3, nous abordons un exemple plus complexe en dimension 4 modélisant le déplacement d'une navette spatiale relativement à une station spatiale. Voir par exemple la figure II.4 qui démontre la capacité de l'approche à calculer avec précision des minorants de temps minimaux d'atteignabilité

- finalement, la sous-section II.4.4 présente un exemple plus abstrait étudiant l'applicabilité de la méthode sur des systèmes de contrôle de plus grande dimension.

Chaque résultat présenté dans ces sous-sections a été démontré rigoureusement à l'aide d'arithmétique d'intervalles, et les certificats duaux de non-atteignabilité p_f ainsi que les valeurs de $J(p_f)$ sont fournis.

0.5.6 Preuves assistées par ordinateur de non-atteignabilité pour des systèmes de contrôle paraboliques

Le chapitre III présente un article co-écrit avec Camille Pouchol, Yannick Privat et Christophe Zhang, en cours de finalisation pour une soumission à un journal scientifique.

Dans ce chapitre, nous étendons la méthode pour étudier la non-atteignabilité de systèmes contrôlés paraboliques de la forme

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{pour presque tout } t \in [0,T], \\ y(0) = y_0 \in X, & \\ u(t) \in \mathcal{U} & \text{pour presque tout } t \in [0,T], \end{cases} \tag{S}$$

où $V \subset X \subset V'$ est un triplet de Gelfand, avec V et X des espaces de Hilbert de dimension infinie. $A : V \to V'$ est un opérateur de domaine $\mathcal{D}(A) = \{x \in V, Ax \in X\}$. Dénotons U un autre espace de Hilbert, avec $B : U \to X$ un opérateur borné. Ici encore, $\mathcal{U} \subset U$ est non vide, convexe, fermé et borné (de borne $M > 0$), et les ensembles cibles $\mathcal{Y}_f$ seront non vides, fermés et bornés.

Dans l'esprit de la méthode en dimension finie, nous voulons utiliser (0.5.26) pour prouver la non-atteignabilité d'ensembles cibles et calculer des minorations de temps minimaux d'atteignabilité, dans une large classe de systèmes contrôlés paraboliques. Toutefois, à notre connaissance aucune preuve assistée par ordinateur d'analyse d'atteignabilité en dimension infinie n'a été considérée dans la littérature.

Le cadre infini-dimensionnel de l'espace d'état complique significativement l'étude de la fonctionnelle de séparation J, requérant sa discrétisation en espace en sus de sa discrétisation en temps. Notamment, les résultats développés dans ce chapitre découlent de multiples théories, telles que la théorie des opérateurs, la théorie des semigroupes et la théorie de l'approximation : voir les sections I.1 et I.3 pour une brève introduction de ces théories.

Dans une première partie (section III.2), nous décrivons une discrétisation de J en temps et espace basée sur la formulation variationnelle de l'équation adjointe (0.5.24). Pour cela, nous supposons la continuité et la coercivité de l'opérateur $-A^*$, et considérons une famille d'espaces de discrétisation $V_h \subset V$ indexée par un paramètre $h > 0$. Nous supposons également que V_h vérifie l'hypothèse standard d'approximation :

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

où $C_0 \geq 0$ est une constante qu'il nous faudra connaître explicitement par la suite. Sur V_h, nous discrétisons A en A_h, et résolvons l'équation adjointe (0.5.24) en temps en la discrétisant avec un schéma d'Euler implicite. À l'aide de théorie de l'approximation et de résultats qualitatifs sur les opérateurs accrétifs, nous obtenons le résultat suivant (voir Proposition III.9) :

$$\|p(t_n) - p_{h,n}\| \leq C_1\|p_f - p_{fh}\| + (C_2h^2 + C_3\Delta t)\|A^*p_f\|, \tag{0.5.31}$$

où $\Delta t > 0$ et $h > 0$ sont les paramètres de discrétisation en temps et espace, et où toutes les constantes sont explicites et dépendant uniquement des constantes de continuité et coercivité de $-A^*$ ainsi que de C_0. En considérant cette discrétisation, on obtient la discrétisation suivante de la fonctionnelle de séparation J :

$$J_{\Delta t,h}(p_{fh}) = \sum_{k=0}^{N_0-1} \sigma_{\mathcal{U}}\left(B^*(\mathrm{Id} - \Delta t A_h)^{-k} p_{fh}\right) + \left\langle y_0, (\mathrm{Id} - \Delta t A_h)^{-N_0} p_{fh}\right\rangle + \sigma_{\mathcal{Y}_f}(p_{fh}). \quad (0.5.32)$$

Par la suite, nous utilisons ce résultat pour en déduire une majoration des erreurs de discrétisation sur J (voir le théorème III.11) :

$$\begin{aligned} |J(p_f) - J_{\Delta t,h}(p_{fh})| \leq \left(\tfrac{1}{2}M\|B\|T\Delta t + (\|y_0\| + M\|B\|T)(C_2h^2 + C_3\Delta t)\right)\|A^*p_f\| \\ + (C_4(\|y_0\| + M\|B\|T) + \|y_f\|)\|p_f - p_{fh}\|. \end{aligned} \quad (0.5.33)$$

Remarquons que dans les deux majorations (0.5.31) et (0.5.33), $p_{fh} \in V_h$ et $p_f \in \mathcal{D}(A^*)$. Comme l'inclusion $V_h \in \mathcal{D}(A^*)$ n'est pas obligatoire, tout $p_{fh} \in V_h$ obtenu par un algorithme de minimisation doit par la suite être interpolé en $p_f \in \mathcal{D}(A^*)$ pour pouvoir appliquer (0.5.33) et certifier la valeur de $J(p_f)$. Par conséquent, la méthodologie décrite dans la sous-section 0.5.5 doit être complétée avec les étapes suivantes :

- la minimisation de $J_{\Delta t,h}$ sur V_h doit être précédée de la preuve de l'estimée ($\mathcal{V}_1$) sur l'espace de discrétisation V_h
- afin de calculer les erreurs de discrétisation finales, il faut interpoler p_{fh} en $p_f \in \mathcal{D}(A^*)$ pour pouvoir vérifier que $J(p_f) < 0$ à l'aide de (0.5.33) et d'arithmétique d'intervalles.

Dans la section (III.3), nous appliquons cette méthodologie pour fournir des preuves de non-atteignabilité sur deux exemples basés sur l'équation de la chaleur. Dans les deux cas, nous discrétisons l'espace d'état à l'aide d'éléments finis $\mathbb{P}_1$, puis interpolons avec des splines cubiques ; voir la section I.3 pour des détails sur ces espaces.

À l'aide de ces résultats, nous étudions d'abord l'équation de la chaleur en une dimension munie de conditions au bord de Dirichlet homogènes. Pour illustrer la précision de la méthode, un cas simple avec contraintes L^2 symétriques et $B = \mathrm{Id}$ est traité. Ensuite, d'autres types de contraintes (L^∞, positives) et restrictions de zones d'actions du contrôle ($B = \chi_\omega$) sont considérées, menant à des preuves de non-atteignabilité et des calculs de minorants de temps minimaux d'atteignabilité pour divers ensembles cibles ; voir par exemple les figures III.2 et III.3.

Ensuite, nous étudions un exemple faisant intervenir deux équations de la chaleur couplées, une seule étant contrôlée par un contrôle soumis à des contraintes L^∞ asymétriques. Nous prouvons l'estimée ($\mathcal{V}_1$), puis prouvons la non-atteignabilité d'une boule centrée en 0 (voir la figure III.4 pour une illustration des conditions initiales et certificat dual de non-atteignabilité). Nous prouvons également la non-atteignabilité d'un ensemble cible non borné de la forme $\{y_f\} \times L^2(0,1)$.

Dans tous les exemples susmentionnés, les preuves assistées par ordinateur utilisent de l'arithmétique d'intervalles rigoureusement codée pour tenir compte des erreurs d'arrondi. Les certificats duaux de non-atteignabilité ainsi que les intervalles contenant les valeurs de J sont fournis et commentés vis-à-vis des divers paramètres. Les erreurs de discrétisation et d'arrondi sont également mentionnées séparément lorsque leur importance relative est remarquable.

0.5.7 Preuves assistées par ordinateur d'atteignabilité pour des systèmes de contrôle de dimension finie

Le chapitre IV est dédié aux preuves assistées par ordinateur d'atteignabilité des systèmes linéaires contrôlés de dimension finie. Il présente des travaux en cours visant à

- prouver l'atteignabilité d'un ensemble cible $\mathcal{Y}_f$
- fournir des approximations de l'ensemble atteignable $\mathcal{R}$, tant extérieures (qui le contiennent) qu'intérieures (qui sont contenues dedans).

Pour développer ces approximations, nous exploitons à nouveau des évaluations de la fonctionnelle de séparation J décrite dans les sous-sections précédentes, ainsi que les discrétisations étudiées dans le chapitre II. Étant donné que chaque évaluation de J prouve la non-atteignabilité d'un demi-espace, l'évaluation de J à plusieurs points distribués de manière relativement homogène autour de l'origine permet l'inclusion de l'ensemble atteignable dans une intersection bornée de demi-espaces : un polytope borné. Augmenter alors le nombre d'évaluations permet d'affiner l'approximation extérieure de l'ensemble atteignable $\mathcal{R}$.

D'autre part, évaluer J permet également de prouver l'atteignabilité de certaines cibles : rappelons que pour chaque évaluation de J, un hyperplan support de $\mathcal{R}$ est découvert, contenant nécessairement un point de $\mathcal{R}$ (qui est supposé convexe, fermé, borné). Il s'ensuit que chaque face de l'approximation extérieure polytopale P de $\mathcal{R}$ contient un point de $\mathcal{R}$, et l'intersection $\mathcal{I}(P)$ de tous les ensembles convexes contenant un point dans chaque face de P est elle-même incluse dans $\mathcal{R}$. Si J a été évaluée à suffisamment de points, $\mathcal{I}(P)$ est non vide (voir Figure IV.2), et augmenter le nombre d'évaluations augmente le volume de cette approximation intérieure. La question de comment choisir optimalement les évaluations de J pour "guider" l'expansion de $\mathcal{I}(P)$ est également considérée : cela permettrait notamment le développement de méthodes pour montrer l'atteignabilité d'un ensemble cible.

Une autre méthode permettant le calcul d'approximations intérieures de $\mathcal{R}$ requiert des informations supplémentaires : rappelons que le calcul de J nécessite le calcul de la fonction support des contraintes. Ce calcul est généralement effectué à l'aide d'une formule explicite du contrôle maximisant le produit scalaire. Il est alors possible de calculer l'état atteint par le système dirigé par ce contrôle, et ainsi d'en déduire une approximation intérieure définie par l'enveloppe convexe de tous ces points. Cette approximation intérieure est alors plus précise que celle précédemment mentionnée (voir Figure IV.4). Tout comme précédemment, la question du choix optimal des évaluations de J est discutée.

Toutes ces approximations de l'ensemble atteignable nécessitent évidemment un usage intensif de preuves assistées par ordinateur, qui s'accompagnent d'une réduction (resp. augmentation) drastique de la taille des approximations intérieures (resp. extérieures) : voir les figures IV.8 et IV.9. D'autre part, de nombreux algorithmes nécessiteraient une adaptation à l'arithmétique d'intervalles (ou arithmétique affine) : par exemple, pour calculer les approximations tant intérieures qu'extérieures, il est souvent nécessaire de passer de représentations des polytopes comme intersections de demi-espaces à des représentations comme enveloppe convexe des sommets. Ces algorithmes sont complexes par nature, et à notre connaissance leur adaptation à l'arithmétique d'intervalles reste à faire. Le chapitre IV ne vise aucunement à présenter une étude approfondie et complète de ce sujet, mais plutôt à introduire les problèmes principaux et des pistes de résolution.

0.5.8 Perspectives et plan de la thèse

Non-atteignabilité en dimension finie. Le chapitre II introduit la méthode pour des systèmes linéaires de dimension finie. L'article traite simultanément une grande variété de systèmes, mais de nombreux autres restent à être traités. Par exemple, nous pensons qu'il serait aisé d'étendre notre méthode à des systèmes non-autonomes non-homogènes $y'(t) = A(t)y(t)+B(t)u(t)+v(t)$, au prix d'erreurs de discrétisation plus élevées. De fait, des contraintes sur le contrôle dépendantes du temps pourraient être considérées de manière similaire, tout comme des contraintes dont la fonction support est inconnue mais approximable (cette approximation devrait alors être à son tour quantifiée explicitement). Une autre piste serait d'explorer des régularisations de la fonctionnelle de séparation, permettant d'obtenir des schémas de discrétisation en temps d'ordre supérieur.

L'étape suivante serait d'étendre la méthode à des systèmes non-linéaires : telle quelle, la méthode dépend fortement sur la linéarité de L_T et la convexité de l'ensemble atteignable ; une telle extension nécessiterait donc de sévères changements. Toujours dans l'esprit d'une fonctionnelle de séparation, d'autres surfaces séparatrices pourraient être considérées, et adaptées à chaque système non-linéaire (rappelons que pour l'instant seuls des hyperplans étaient considérés) ; ces surfaces pourraient être déterminées par exemple à l'aide d'outils venant du contrôle optimal, par exemple les caractérisations des équations adjointes comme le Principe du Maximum de Pontryagin. Une autre piste serait de considérer l'EDP de transport linéaire associée au système via l'opérateur de Koopman, et appliquer la méthode à l'EDP linéaire d'une manière similaire à celle employée dans le chapitre III.

Non-atteignabilité en dimension infinie. Comme vu précédemment, les preuves assistées par ordinateur pour la non-atteignabilité de systèmes contrôlés en dimension infinie sont rendues significativement plus complexes par la nécessité d'une discrétisation en espace en complément de celle en temps. En particulier, nous n'avons étudié que des équations paraboliques avec des opérateurs continus et coercifs, donc une première piste d'extension serait de considérer d'autres opérateurs. Tout comme en dimension finie, le point essentiel de la méthodologie est l'obtention de majorations d'erreurs explicites et petites : une extension aux équations hyperboliques requerrait donc une analyse minutieuse des schémas de discrétisation pour éviter une explosion des erreurs de discrétisation avec le temps final. Un premier objectif serait donc de s'attaquer à d'autres systèmes linéaires paraboliques, en incluant par exemple des dépendances en temps ou des conditions aux bords plus complexes. Une étude de contrôle au bord serait également intéressante, mais est significativement compliquée par le caractère non borné de B, jusqu'alors nécessaire pour rendre explicites les erreurs de discrétisation.

Atteignabilité en dimension finie. Tel qu'exposé précédemment, un autre objectif des recherches à venir est le développement d'une méthodologie pour certifier numériquement l'atteignabilité d'un ensemble cible en dimension finie. Ceci est d'autant plus complexe qu'il requiert le calcul d'approximations extérieures de l'ensemble atteignable, ce qui est très coûteux et particulièrement avec de l'arithmétique d'intervalles. Un but similaire encore plus coûteux est le calcul d'approximations intérieures de l'ensemble atteignable. Comme ces approximations ne dépendent que des évaluations de la fonctionnelle J, les mêmes extensions que pour la non-atteignabilité en dimension finie pourraient par la suite être considérées. En revanche, les cas non linéaires ou les systèmes en dimension infinie semblent à ce stade inenvisageables : les premiers du fait de la non-convexité de l'ensemble atteignable, et les seconds car seuls des sous-ensembles de dimension finie d'un ensemble atteignable de dimension infinie pourraient être calculés.

Plan de la thèse. Cette thèse est divisée en quatre chapitre successifs. Le chapitre I introduit de nombreuses théories et outils mathématiques essentiels pour les parties principales de la thèse : théorie des opérateurs et des semigroupes, analyse convexe, théorie de l'approximation et calculs numériques certifiés seront ainsi successivement introduits. Le chapitre II formalise alors la méthode décrite dans l'introduction et l'applique à la non-atteignabilité des systèmes de dimension finie. Ensuite, le chapitre III étend cette méthode pour aborder la non-atteignabilité des systèmes linéaires paraboliques. Finalement, le chapitre IV présente des travaux en cours visant à formaliser des preuves assistées par ordinateur d'atteignabilité pour des systèmes linéaires contrôlés de dimension finie.

I

Tools

Contents

In this chapter, we will introduce a few essential tools to develop the computer-assisted proofs in the following chapters. Each section is independent, even though some choices will be justified with a reference to other sections. Note in particular that some notations may be reused from one section to another while changing of properties (in particular, $\mathcal{A}$ will denote a coercive operator in section I.3, and its opposite in section I.1).

Firstly, Section I.1 will introduce some results of operator theory, along with estimates bounding the norms of functions of operators. These will prove determinant for the upper-bounding of discretisation errors of the partial differential equations in Chapter III, as well as some properties in the finite-dimensional part.

Some notions of semigroup theory will also be presented: definitions and characterisations of C_0 semigroups of contraction will be recalled. These properties are omnipresent in the infinite-dimensional part of the thesis (Chapter III), and underlie many properties of the rest of the thesis.

Next, Section I.2 we present standard definitions and properties of convex functionals. We give details about the separation argument central to the whole non-reachability approach of the thesis, as well as algorithms used to solve primal-dual problems. Regularisations available to increase their rates of convergence and obtain nice properties of the minimisers are also considered.

In Section I.3, we introduce classical discretisation and interpolation spaces, and discuss their advantages and drawbacks. Some known estimates will be proved, paying attention to the tracking and optimisation of constants, whose knowledge and smallness is fundamental to all computer-assisted proofs based on discretisations.

Finally, in Section I.4, we will present the rigorous numerical methods used to tackle the round-off errors omnipresent in every numerical computation. In particular, different techniques minimising the so-called *wrapping effect* as well as the computation time of the functional J central to our approach will be introduced.

All along this chapter, X will denote a Hilbert space endowed with inner product $\langle \cdot, \cdot \rangle_X$ and norm $\| \cdot \|_X$. All our assumptions are concerned with a complex Hilbert space X, while our separation argument is stated within a real Hilbert space, also denoted by X. This is because all our (non-)reachability results will be about initial conditions, targets, equations which are in $\mathbb{R}$, while discussing fine properties of operators requires the complex world.

Hence, with a slight abuse of notation, the state space X will stand either for $\mathbb{R}^n$ or $\mathbb{C}^n$ in the finite-dimensional case, either for real-valued L^2 functions or complex-valued functions in the infinite-dimensional case. Of course, we could instead have stated our separation argument in the complex Hilbert space setting, but at the expense of keeping the real part of all inner products throughout.

The topological dual $X' = \mathcal{L}(X, \mathbb{C})$ (or $\mathcal{L}(X, \mathbb{R})$) will be endowed with its natural norm $\| \cdot \|_{\mathcal{L}(X,\mathbb{C})}$ (respectively $\| \cdot \|_{\mathcal{L}(X,\mathbb{R})}$). All these subscripts may be dropped if no confusion is possible.

$\mathcal{A}$ will generally denote a linear operator defined on a subspace of $\mathcal{D}(\mathcal{A}) \subset X$, but its precise definition and properties will vary from section I.1 ($\mathcal{A} : \mathcal{D}(\mathcal{A}) \to X$ generator of a C_0 semigroup of contractions) to section I.3 (coercive and continuous operator $\mathcal{A} : V \subset X \to V'$).

I.1 Semigroups and operator theory

In this first section, we will introduce tools and recall results from operator theory and semigroup theory. Even though X will denote a Hilbert space, some of the results can be applied in Banach spaces. These results will prove useful in Chapter III for the study of discretisation errors.

I.1.1 Operator theory

In this section we introduce a few elements of operator theory. Proofs and details can be found, for example, in [BCL99] or [Haa06]. We first recall basic definitions.

Definition I.1. Consider an unbounded linear operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset X \to X$. $\mathcal{A}$ is said to be bounded (or continuous) if $\mathcal{D}(\mathcal{A}) = X$ and if there exists $c > 0$ such that

$$\|\mathcal{A}x\| \leq c\|x\| \quad \forall\, x \in \mathcal{D}(\mathcal{A}),$$

and closed if for all convergent sequence $x_n \to x$ in $\mathcal{D}(\mathcal{A})$ such that $\mathcal{A}x_n$ converges to y, then $x \in \mathcal{D}(\mathcal{A})$ and $y = \mathcal{A}x$. Furthermore, we define the resolvent set, the spectrum and the numerical range of $\mathcal{A}$ as

$$\rho(\mathcal{A}) = \left\{ \lambda \in \mathbb{C}, \quad (\lambda \operatorname{Id} - \mathcal{A})^{-1} \text{ is a bounded linear operator in } X \right\} \tag{I.1.1}$$

$$\sigma(\mathcal{A}) = \mathbb{C} \setminus \rho(\mathcal{A}) \tag{I.1.2}$$

$$W(\mathcal{A}) = \left\{ \langle \mathcal{A}x, x \rangle, \quad x \in \mathcal{D}(\mathcal{A}), \|x\| = 1 \right\}. \tag{I.1.3}$$

A few interesting properties of those sets are the following:

Theorem I.2. Let $\mathcal{A}$ be a closed unbounded operator. Its numerical range $W(\mathcal{A})$ is convex, and for all $\lambda \in \rho(\mathcal{A})\backslash\overline{W(\mathcal{A})}$,

$$\|(\lambda \operatorname{Id} - \mathcal{A})^{-1}\| \leq \frac{1}{\operatorname{dist}(\lambda, \overline{W(\mathcal{A})})}. \tag{I.1.4}$$

If furthermore $\mathcal{A}$ is bounded, then

$$\sigma(\mathcal{A}) \subset \overline{W(\mathcal{A})}. \tag{I.1.5}$$

We now introduce sectorial operators, and functions of those operators, leading to Theorem I.7, which will prove crucial to bound discretisation errors in Chapter III.

Definition I.3. Let $\alpha \in [0, \frac{\pi}{2}]$. We say that $\mathcal{A}$ is mα-accretive, or sectorial with angle α if

$$W(\mathcal{A}) \subset \mathcal{S}_\alpha := \{z \in \mathbb{C}, |\arg z| \leq \alpha\}, \tag{I.1.6}$$

and

$$\forall\, z \notin \mathcal{S}_\alpha, \quad z \operatorname{Id} - \mathcal{A} \text{ is an isomorphism from } \mathcal{D}(\mathcal{A}) \text{ to } X. \tag{I.1.7}$$

If $\alpha = \frac{\pi}{2}$, $\mathcal{A}$ is said to be m-accretive.

This definition allows for the definition of rational functions, and in turn holomorphic functions, of operators. Note that weaker definitions have been considered in the literature, we refer to [Haa06] for bountiful details.

Remark I.4. Notice that in our Hilbert framework, (I.1.7) follows automatically from (I.1.6) using Theorem I.2. Notice as well that continuous and coercive operators are always sectorial – we will take advantage of this property in Chapter III.

In the following propositions, we introduce functions of sectorial operators.

Proposition I.5. Assume that $\mathcal{A}$ is mα-accretive, with $\alpha \in [0, \frac{\pi}{2}]$. Let $n \in \mathbb{N}$, and $(\alpha_j)_{j\in\{1,\dots,n\}}$ a sequence of complex numbers in $\mathbb{C}\backslash\mathcal{S}_\alpha \subset \rho(\mathcal{A})$. Let as well $(m_j) \in (\mathbb{N}^*)^n$, and $(r_j)_{j\in\{1,\dots,n\}}, r(\infty) \in \mathbb{C}$. Define the rational function

$$r : \begin{cases} \mathbb{C} & \to \mathbb{C} \\ z & \mapsto r(\infty) + \sum_{j=1}^n \frac{r_j}{(\alpha_j - z)^{m_j}}. \end{cases} \tag{I.1.8}$$

One can then define the operator $r(\mathcal{A})$ as

$$r(\mathcal{A}) = r(\infty) \operatorname{Id} + \sum_{j=1}^n r_j(\alpha_j \operatorname{Id} - \mathcal{A})^{-m_j}, \tag{I.1.9}$$

which has domain $\mathcal{D}(r(\mathcal{A})) = X$.

Proposition I.6. For $\alpha \in [0, \frac{\pi}{2}]$, let f a holomorphic function in $\mathcal{S}_\alpha$, continuous and bounded on $\overline{\mathcal{S}_\alpha}$, such that f is the limit of a uniformly convergent sequence of rational functions $(r_n)_{n\in\mathbb{N}}$ with poles in $\mathbb{C} \setminus \overline{\mathcal{S}_\alpha}$, that is,

$$\lim_{n\to+\infty} \sup_{z\in\mathcal{S}_\alpha} |f(z) - r_n(z)| = 0. \tag{I.1.10}$$

$(r_n(\mathcal{A}))_{n\in\mathbb{N}}$ then converges in operator norm towards a limit that we define as $f(\mathcal{A})$.

Note that this proposition is an alternative to the formal definitions of exponentials $t \mapsto \exp(-t\mathcal{A})$ of mα-accretive operators $\mathcal{A}$ – in other words, semigroups.

Theorem I.7 (Crouzeix-Delyon [CD03])**.** For $\alpha \in [0, \frac{\pi}{2}]$, let f be a holomorphic function in $\mathcal{S}_\alpha$, continuous and bounded on $\overline{\mathcal{S}_\alpha}$. The following estimate holds

$$\|f(\mathcal{A})\|_{\mathcal{L}(X)} \leq C_\alpha \sup_{z\in\mathcal{S}_\alpha} |f(z)|, \tag{I.1.11}$$

where

$$\frac{\pi}{2} \leq C_0 \leq 2 + \frac{2}{\sqrt{3}} \quad \text{and} \quad \frac{\pi \sin\alpha}{2\alpha} \leq C_\alpha \leq \min\left(\frac{\pi-\alpha}{\alpha}, C_0\right). \tag{I.1.12}$$

This result is one of the numerous results estimating the norms of functions of operators. See also [Cro07; CP17; CG19] and the references therein for different cases of bounded / unbounded operators, which numerical ranges are included in various complex domains.

I.1.2 Semigroup theory

Semigroups are the infinite-dimensional adaptation of the matrix exponentials in the sense that they naturally solve equations of the type $y' = \mathcal{A}y$. Notice that for $\mathcal{A}$ a given sectorial operator, Proposition I.6 guarantees that the family of operators $(\exp(-t\mathcal{A}))_{t\geq 0}$ is well defined on X. The theory of semigroups of linear operators, that we will study in this section, extends this definition and analyses the properties of such functions. We do not claim to offer a comprehensive review of the field, and refer to classical references such as [Paz12; EN06]. For more general results of functional analysis, one can consult [BCL99; Haa06].

Definition I.8. A family $(T(t))_{t\geq 0}$ of operators from X to X is a semigroup if

- $T(0) = \mathrm{Id}$,
- $\forall\, t, s \geq 0, T(t+s) = T(t)T(s)$.

Furthermore, it is said to be strongly continuous, or simply C_0, if

$$\forall\, x \in X, \quad \lim_{t\to 0^+} T(t)x = x. \tag{I.1.13}$$

Finally, it is said to be a semigroup of contractions if

$$\forall\, t \geq 0, \quad \|T(t)\| \leq 1. \tag{I.1.14}$$

Definition I.9. Given $(T(t))_{t\geq 0}$ a semigroup, the linear operator $\mathcal{A}$ defined by

$$\mathcal{D}(\mathcal{A}) = \left\{ x \in X, \quad \lim_{t\to 0^+} \frac{T(t)x - x}{t} \text{ exists} \right\} \tag{I.1.15}$$

and

$$\forall\, x \in \mathcal{D}(\mathcal{A}), \quad \mathcal{A}x = \lim_{t\to 0^+} \frac{T(t)x - x}{t} \tag{I.1.16}$$

is called the infinitesimal generator of the semigroup $(T(t))_{t\geq 0}$.

Theorem I.10 (Hille-Yosida)**.** An operator $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ in a Hilbert space X is the infinitesimal generator of a C_0 semigroup of contractions $(T(t))_{t\geq 0}$ if and only if

1. $\mathbb{R}_+^* \subset \rho(\mathcal{A})$
2. $\forall\, \lambda > 0,\ \|(\mathcal{A} - \lambda\,\mathrm{Id})^{-1}\| \leq \frac{1}{\lambda}$.

Remark I.11. Note that if $\mathcal{A}$ satisfies that $-\mathcal{A}$ is mα-accretive for $\alpha \in [0, \frac{\pi}{2}]$, then it is easy to check using Theorem I.2 that the conditions of Hille-Yosida's theorem hold and thus prove that $\mathcal{A}$ generates a C_0 semigroup of contractions.

Using the following definition and Theorem I.13, one can link semigroups of contractions with sectorial operators.

Definition I.12. We say that $\mathcal{A}$ is dissipative if for all $x \in \mathcal{D}(\mathcal{A})$, $\mathrm{Re}\,\langle \mathcal{A}x, x\rangle \leq 0$. In other words, that $-\mathcal{A}$ is m-accretive.

Theorem I.13 (Lumer-Phillips)**.** Let X be a Hilbert space, and $\mathcal{A}$ be an operator defined on $\mathcal{D}(\mathcal{A})$, which is included and dense in X. We have the following equivalence:

- $\mathcal{A}$ is dissipative and there exists $\lambda_0 > 0$ such that $\lambda_0 \in \rho(\mathcal{A})$
- $\mathcal{A}$ is closed, and both $\mathcal{A}$ and $\mathcal{A}^*$ are dissipative
- $\mathcal{A}$ generates a C_0 semigroup of contractions on X.

Note that both those theorems can be applied to obtain more general bounds on semigroups: where semigroups of contractions "only" require a general bound by 1, it is sometimes preferred to obtain bounds of the type $\exp(-\omega t)$ for $\omega > 0$. This can easily be obtained using the previous theorems by shifting the operators (and the hypotheses) by $\omega\,\mathrm{Id}$.

I.2 Convex Analytic Tools

Many core components of this thesis rely on convex analysis. Therefore, this section will introduce its most crucial elements. In all this section (and this section only), we consider that X is a real-Hilbert space (see Page 32 for details). All results would still hold in the complex setting, at the cost of adding a real part when needed. We shall also identify X and its dual X', and, unless specified otherwise, the topology considered for X will be the weak topology.

Most results will not be proved, we refer to standard textbooks, for example [Roc70], [Rud91], [BC11] or [Sch13].

I.2.1 Preliminary definitions

Let $f : X \to \overline{\mathbb{R}}$.

Definition I.14. Recall the standard definitions:

- f is proper if there exists $x \in X$ such that $f(x) < +\infty$ and if every $x \in X$ satisfies $f(x) > -\infty$
- f is convex if for all $x, y \in X$, for all $\lambda \in [0, 1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$
- f is lower semi-continuous if for all $x_0 \in X$, $\liminf\limits_{x\to x_0} f(x) \geq f(x_0)$.

We define $\Gamma_0(X)$ as the functional space of proper, convex and lower semi-continuous functions from X to $\overline{\mathbb{R}}$.

From now on we will assume $f \in \Gamma_0(X)$. We define the convex conjugate of f as

$$f^* : \begin{cases} X & \to \overline{\mathbb{R}} \\ y & \mapsto \sup\limits_{x \in X} \langle y, x \rangle - f(x). \end{cases} \tag{I.2.1}$$

Theorem I.15 (Fenchel-Moreau)**.** Let $f \in \Gamma_0(X)$. Then $f^* \in \Gamma_0(X)$ and $(f^*)^* = f$.

Two kinds of convex functional will be especially useful in this thesis:

Definition I.16. Let $C \subset X$. Define the indicator of C as

$$\delta_C : \begin{cases} X & \to \overline{\mathbb{R}} \\ x & \mapsto \begin{cases} 0 & \text{if} \quad x \in C \\ +\infty & \text{if} \quad x \notin C, \end{cases} \end{cases} \tag{I.2.2}$$

and its support function as:

$$\sigma_C = \delta_C^* : \begin{cases} X & \to \overline{\mathbb{R}} \\ y & \mapsto \sup\limits_{x \in C} \langle y, x \rangle. \end{cases} \tag{I.2.3}$$

Proposition I.17. We have the equivalence:

$$\delta_C \in \Gamma_0(X) \iff C \text{ is nonempty, convex and closed.} \tag{I.2.4}$$

Proposition I.18. Let $C \subset X$ be nonempty, closed, convex and bounded. Then σ_C is a M-Lipschitz functional, where $M = \sup_{x \in C} \|x\|$.

Proof. It is easily seen that

$$\forall\, x, y \in X, \quad \sigma_C(x) \leq \sigma_C(x - y) + \sigma_C(y) \leq M\|x - y\| + \sigma_C(y), \tag{I.2.5}$$

which clearly implies that

$$\forall\, x, y \in X, \quad \sigma_C(x) - \sigma_C(y) \leq M\|x - y\|, \tag{I.2.6}$$

and a symmetry argument concludes that σ_C is M-Lipschitz. □

The following theorem, which is a direct consequence of Theorem I.20 when $A = \{x\}$, characterises closed convex sets:

Theorem I.19. Let $C \subset X$ be closed, convex. Then, for $x \in X$,

$$x \in C \iff \forall\, y \in X, \quad \langle y, x \rangle \leq \sigma_C(y). \tag{I.2.7}$$

A similar result, whose proof can be found in [Rud91, Theorem 3.4] is the following:

Theorem I.20 (Hahn-Banach)**.** Suppose A and B are disjoint, nonempty, convex sets. If

A is (weakly) compact and B is closed, there exist $p_f \in X$ such that

$$\sup_{a\in A}\langle p_f, a\rangle < \inf_{b\in B}\langle p_f, b\rangle. \tag{I.2.8}$$

Note that the converse is true: existence of a strictly separating hyperplane guarantees that the two sets are disjoint. Using the vocabulary of this thesis, the Hahn-Banach Theorem can obviously be rewritten as follows: suppose that A and B are nonempty convex sets in a real-valued Hilbert space X, and assume that A is weakly compact and B closed. Then

$$A \cap B = \emptyset \iff \exists\, p_f \in X,\ \sigma_A(p_f) + \sigma_B(-p_f) < 0. \tag{I.2.9}$$

The converse implication is the cornerstone of the computer-assisted proofs of non-reachability in this thesis, while the direct one guarantees the sharpness of our approach. In the following theorem, we present a more precise characterisation, along with a proof.

Theorem I.21. Suppose A and B are nonempty, convex sets such that A is weakly compact and B is closed. Then there exist $x^\star \in A$ and $y^\star \in B$ such that

$$\|x^\star - y^\star\| = \inf_{(x,y)\in A\times B}\|x-y\|. \tag{I.2.10}$$

Denoting $p^\star = y^\star - x^\star$, we then have that

$$\sigma_A(p^\star) + \sigma_B(-p^\star) = -\|p^\star\|^2 = \|p^\star\| \inf_{\|p_f\|\le 1} \sigma_A(p_f) + \sigma_B(-p_f). \tag{I.2.11}$$

Finally, the following statements are equivalent:

1. $A \cap B \neq \emptyset$
2. $x^\star = y^\star$
3. $\sigma_A(p^\star) + \sigma_B(-p^\star) \ge 0$
4. $\forall\, p_f \in X,\ \sigma_A(p_f) + \sigma_B(-p_f) \ge 0.$

Remark I.22. It follows from this theorem that

$$\operatorname{dist}(A,B) = -\inf_{\|p_f\|\le 1} \sigma_A(p_f) + \sigma_B(-p_f), \tag{I.2.12}$$

which gives another intuition of why this is true.

Proof. Consider (x_n, y_n) a minimising sequence of (I.2.10). Then, by weak-compactness of A, (x_n) converges weakly, up to a subsequence, to a $x^\star \in A$. Let us denote $y^\star = \Pi_B x^\star$. Since the norm is weakly lower semicontinuous, it follows that

$$\inf_{(x,y)\in A\times B}\|x-y\| = \inf_{y\in B}\|x^\star - y\| = \|x^\star - y^\star\| = \inf_{x\in A}\|x - y^\star\|, \tag{I.2.13}$$

and thus $x^\star = \Pi_A y^\star$ and the pair $(x^\star, y^\star)$ is optimal.

Let us now prove (I.2.11). Recall the classical characterisation of the projection on a closed convex set:

$$x^\star = \Pi_A y^\star \iff \forall\, x \in A,\quad \langle x - x^\star, y^\star - x^\star\rangle \le 0 = \langle x^\star - x^\star, y^\star - x^\star\rangle. \tag{I.2.14}$$

It follows that

$$\sigma_A(p^\star) = \sup_{x\in A}\langle x, y^\star - x^\star\rangle = \sup_{x\in A}\langle x - x^\star, y^\star - x^\star\rangle + \langle x^\star, y^\star - x^\star\rangle = \langle x^\star, p^\star\rangle. \tag{I.2.15}$$

Similarly,

$$\sigma_B(-p^\star) = -\langle y^\star, p^\star\rangle. \tag{I.2.16}$$

This entails that $\sigma_A(p^\star) + \sigma_B(-p^\star) = -\|p^\star\|^2$. To prove the right-hand equality of (I.2.11), notice that for any $p_f \in X$

$$\begin{aligned}\sigma_A(p_f) + \sigma_B(-p_f) &= \sup_{x\in A}\langle x, p_f\rangle + \sup_{y\in B}\langle y, -p_f\rangle \\ &\geq \sup_{(x,y)\in A\times B} \langle x, p_f\rangle - \langle y, p_f\rangle = \sup_{(x,y)\in A\times B} \langle x-y, p_f\rangle \\ &\geq \sup_{(x,y)\in A\times B} -\|x-y\|\|p_f\| = -\|p_f\| \inf_{(x,y)\in A\times B} \|x-y\| \\ &\geq -\|p_f\|\|p^\star\|.\end{aligned}$$

Let us now prove the equivalences: (1) and (2) are evidently equivalent by definition of the projection. The equivalence of (2), (3) and (4) follows immediately from (I.2.11) and from the 1-homogeneity of support functions. □

In this thesis, we are concerned with the special case happening when B is the image of a convex set through a linear map, particularly relevant in linear control theory. In the following sections, we will therefore focus on this setting, and start by introducing the Fenchel-Rockafellar framework, which provides interesting properties of strong duality, as well as another proof of Theorem I.21.

I.2.2 Fenchel-Rockafellar duality

In this section we consider two Hilbert spaces E and X linked by a bounded linear operator $K : E \to X$. For $f \in \Gamma_0(E)$ and $g \in \Gamma_0(X)$, we study the minimisation problem

$$\inf_{x\in E} f(x) + g(Kx). \tag{I.2.17}$$

Proposition I.23. Introducing the dual minimisation problem

$$\inf_{y\in X} f^*(K^*y) + g^*(-y). \tag{I.2.18}$$

we have a weak duality property linking the two minimisation problems (I.2.17) and (I.2.18):

$$\inf_{x\in E} f(x) + g(Kx) \geq -\inf_{y\in X} f^*(K^*y) + g^*(-y). \tag{I.2.19}$$

Under weak conditions, this inequality becomes an equality. The following theorem was first proved in [Roc67]:

Theorem I.24 (Rockafellar)**.** Assume that

$$\exists\, y \in E,\quad g^*(-y) < +\infty \text{ and } f^* \text{ is continuous at } K^*y. \tag{I.2.20}$$

Then strong duality holds, in the sense that

$$\inf_{x\in E} f(x) + g(Kx) = -\inf_{y\in X} f^*(K^*y) + g^*(-y). \tag{I.2.21}$$

Furthermore, should that infimum be finite, then the primal problem (I.2.17) admits a minimiser.

Remark I.25. The condition (I.2.20) is a sufficient but not necessary condition for the strong duality. See [Roc67] or [BC11] for more general conditions.

It follows from this theorem that, given $\mathcal{E} \subset E$ and $\mathcal{Y} \subset X$ two convex sets, and $K : E \to X$ and assuming that $\mathcal{E}$ is weakly compact and $\mathcal{Y}$ is closed (ensuring that (I.2.20) is verified) the following formulations are equivalent:

- $K\mathcal{E} \cap \mathcal{Y} \neq \emptyset$
- $\inf_{u\in E} \delta_{\mathcal{E}}(u) + \delta_{\mathcal{Y}}(Ku) = -\inf_{p_f \in X} \sigma_{K\mathcal{E}}(p_f) + \sigma_{\mathcal{Y}}(-p_f) = 0.$

Under those conditions, one can wonder how to find the minimisers of these two minimisation problems. This is the objective of the following section.

I.2.3 Saddle-point algorithms

The combination of the two problems (I.2.17) and (I.2.18) can reformulated as the following generic saddle-point problem

$$\inf_{x\in E} \sup_{y\in X} \langle Kx, y\rangle + f(x) - g^*(y). \tag{I.2.22}$$

This formulation has been extensively studied in the literature, but we shall here only present one widely used primal-dual algorithm, presented in [CP11], whose content is summarised in Algorithm I.1.

Algorithm I.1: Chambolle-Pock algorithm [CP11]

```
**Data:** $\tau, \sigma > 0$ with $\tau\sigma\|K\|^2 \leq 1$,  $n \in \mathbb{N}$,  $(x^0, y^0) \in E \times X$
**Result:** $x^n$, $y^n$
$\overline{x}^0 = x^0$;
$k = 0$;
while $k < n$ do
    $y^{k+1} = -\operatorname{prox}_{\sigma g^*}(-y^k + \sigma K\overline{x}^k)$ ;
    $x^{k+1} = \operatorname{prox}_{\tau f}(x^k + \tau K^* y^{k+1})$ ;
    $\overline{x}^{k+1} = 2x^{k+1} - x^k$;
```

This algorithm makes use of proximal operators, defined as follows: for $f \in \Gamma_0(X)$,

$$\operatorname{prox}_f : \left\{ \begin{array}{ccc} X & \longrightarrow & X \\ x & \longmapsto & \operatorname{argmin}_{y\in X} \left\{ \frac{1}{2}\|x-y\|^2 + f(y) \right\}. \end{array} \right. \tag{I.2.23}$$

This function is well defined because the functional minimised is strongly convex and thus admits exactly one minimiser.

One can either compute the proximal operator of f or of its Fenchel conjugate, thanks to Moreau's identity:

$$\forall\, x \in E, \quad x = \operatorname{prox}_{\tau f}(x) + \tau \operatorname{prox}_{\tau^{-1} f^*}(\frac{x}{\tau}). \tag{I.2.24}$$

In particular, one can easily compute the proximal operators associated to indicator and support functions:

Proposition I.26. Let C be a convex and closed subset of X. Denoting Π_C the orthogonal projection on C, we have

$$\forall\, x \in X, \quad \operatorname{prox}_{\delta_C}(x) = \Pi_C(x) \quad \text{and} \quad \operatorname{prox}_{\sigma_C}(x) = x - \Pi_C(x). \tag{I.2.25}$$

Under the assumption of the existence of a saddle-point of (I.2.22), and assuming that $\tau\sigma\|K\|^2 < 1$ and $\theta = 1$, the Chambolle-Pock algorithm I.1 has guaranteed convergence towards a saddle-point – see [CP11] for details. When considering our problem of intersection $K\mathcal{E} \cap \mathcal{Y}$, it just so happens that the existence of a saddle-point is equivalent to $K\mathcal{E} \cap \mathcal{Y} \neq \emptyset$.

Thankfully, (I.2.9) entails that we do not need the convergence to solve the question of emptiness of $K\mathcal{E} \cap \mathcal{Y}$ – merely existence of a negative point. Nevertheless, we study in the next section a regularisation of this problem that allows us to obtain convergence guarantees as well as interesting properties of the dual minimiser.

I.2.4 Regularisation of the primal-dual problem

We shall consider here a family of regularisations of the dual problem (I.2.18) enhancing the convexity of the functional, while preserving the sign property (I.2.9).

Proposition I.27. Assume that $\mathcal{E}$ is convex and weakly compact, and $\mathcal{Y}$ is convex, closed and bounded. Given a closed operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \to \overline{\mathcal{D}(\mathcal{A})} = X$ and a $\lambda > 0$, we have the equivalence

$$\exists p_f \in X,\ \sigma_{K\mathcal{E}}(p_f)+\sigma_{\mathcal{Y}}(-p_f) < 0 \iff \exists p_f \in \mathcal{D}(\mathcal{A}),\ \sigma_{K\mathcal{E}}(p_f)+\sigma_{\mathcal{Y}}(-p_f)+\frac{\lambda}{2}\|\mathcal{A}p_f\|^2 < 0. \tag{I.2.26}$$

Furthermore, if there exists $\kappa > 0$ such that for all $p_f \in \mathcal{D}(\mathcal{A})$, $\|\mathcal{A}p_f\| \geq \kappa\|p_f\|$, the following minimisation problem admits a unique minimiser $p_f^\star \in \mathcal{D}(\mathcal{A})$:

$$\inf_{p_f \in \mathcal{D}(\mathcal{A})} \sigma_{K\mathcal{E}}(p_f) + \sigma_{\mathcal{Y}}(-p_f) + \frac{\lambda}{2}\|\mathcal{A}p_f\|^2. \tag{I.2.27}$$

Proof. The equivalence (I.2.26) follows easily from the 1-homogeneity of support functions: since the reverse implication is obvious, assume the left-hand side, that is, the existence of $p_f \in X$ such that $\sigma_{K\mathcal{E}}(p_f) + \sigma_{\mathcal{Y}}(-p_f) < 0$. Then since both $K\mathcal{E}$ and $\mathcal{Y}$ are bounded, Proposition I.18 guarantees that $\sigma_{K\mathcal{E}}+\sigma_{\mathcal{Y}}$ is Lipschitz continuous, and by density of $\mathcal{D}(\mathcal{A})$, there exists $p_f^\star \in \mathcal{D}(\mathcal{A})$ such that $\sigma_{K\mathcal{E}}(p_f^\star) + \sigma_{\mathcal{Y}}(-p_f^\star) < 0$. Taking $\mu > 0$, we finally have that

$$\frac{1}{\mu}\left(\sigma_{K\mathcal{E}}(\mu p_f^\star) + \sigma_{\mathcal{Y}}(-\mu p_f^\star) + \frac{\lambda}{2}\|\mu\mathcal{A}p_f^\star\|^2\right) = \sigma_{K\mathcal{E}}(p_f^\star) + \sigma_{\mathcal{Y}}(-p_f^\star) + \frac{\mu\lambda}{2}\|\mathcal{A}p_f^\star\|^2, \tag{I.2.28}$$

which converges to $\sigma_{K\mathcal{E}}(p_f^\star) + \sigma_{\mathcal{Y}}(-p_f^\star) < 0$ as μ converges to 0. This concludes the existence of $q \in \mathcal{D}(\mathcal{A})$ such that

$$\sigma_{K\mathcal{E}}(p_f) + \sigma_{\mathcal{Y}}(-p_f) + \frac{\lambda}{2}\|\mathcal{A}p_f\|^2 < 0. \tag{I.2.29}$$

This functional's restriction on $\mathcal{D}(\mathcal{A})$ being continuous (because $K\mathcal{E}$ and $\mathcal{Y}$ are bounded), and strongly convex by assumption, the existence and uniqueness of the minimiser of (I.2.27) follows. □

This regularisation favours dual variable small with respect to high singular values of $\mathcal{A}$, which can prove essential. To apply the algorithm mentioned in the last section, one can compute the appropriate primal problem (in the sense of Fenchel-Rockafellar duality), which satisfies the hypotheses of Rockafellar's Theorem I.24:

$$\inf_{u \in E} \delta_{\mathcal{E}}(u) + \frac{1}{2\lambda}\operatorname{dist}_{\|\mathcal{A}^{-1}\cdot\|}(L_T u, \mathcal{Y})^2, \tag{I.2.30}$$

where $\operatorname{dist}_{\|\mathcal{A}^{-1}\cdot\|}(x, \mathcal{Y}) = \inf_{y \in \mathcal{Y}} \|\mathcal{A}^{-1}(x-y)\| \in \mathbb{R} \cup \{+\infty\}$ for all $x \in X$. The primal problem is also regularised, while preserving the properties of nonnegativity and equality to zero whenever $u \in \mathcal{E}$ and $Ku \in \mathcal{Y}$.

Denoting $\tilde{\sigma}_{\mathcal{Y}} = \sigma_{\mathcal{Y}} + \frac{\lambda}{2}\|\mathcal{A}\cdot\|^2$, we compute the proximal operators of the regularised functionals:

$$\forall p_f \in X, \qquad \operatorname{prox}_{\tilde{\sigma}_{\mathcal{Y}}}(p_f) = (\operatorname{Id} + \lambda\mathcal{A}^*\mathcal{A})^{-1}(p_f - \Pi_{\mathcal{Y}}(p_f)), \tag{I.2.31}$$

and $\operatorname{prox}^*_{\tilde{\sigma}_{\mathcal{Y}}}$ can be computed using Moreau's identity (I.2.24). For this new primal-dual problem, [CP11] proves the convergence of Algorithm I.1, as well as adapted algorithms for uniformly convex functionals guaranteeing better rates of convergence. The regularisation parameter λ

also increases the speed of convergence: in essence, for the adapted version of Chambolle-Pock's algorithm in finite-dimensional spaces E and X, the rate of convergence is of the form $\mathcal{O}(\frac{1}{\lambda N})$, where N is the number of iterations.

Another benefit from regularisation is to obtain properties on the minimiser. In our context, the main result is the following, stating that the minimiser's direction is independent of the regularisation parameter λ. It is a consequence of the following lemma:

Lemma I.28. Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be two convex continuous functions such that

$$\forall\,\alpha \geq 0, \forall\,x \in X, \quad \begin{cases} f(\alpha x) = \alpha f(x) \\ g(\alpha x) = \alpha^2 g(x). \end{cases} \tag{I.2.32}$$

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

$$\inf_{x\in X} f(x) + \frac{\lambda}{2} g(x) = -\frac{1}{2\lambda} \sup_{s\in S_g} (\min(f(s), 0))^2, \tag{I.2.33}$$

and if the sup is reached, then:

$$\exists\, p^\star \in X, \forall\,\lambda > 0, \exists\, r > 0, \quad f(rp^\star) + \frac{\lambda}{2} g(rp^\star) = \inf_{x\in X} f(x) + \frac{\lambda}{2} g(x). \tag{I.2.34}$$

Proof. We have that $\forall\,x \in X$ such that $g(x) < +\infty$, there exists $r > 0, s \in S_g$ such that $x = rs$. Letting $\lambda > 0$, Note now that

$$\begin{aligned} \inf_{x\in X} f(x) + \frac{\lambda}{2} g(x) &= \inf_{\substack{x\in X \\ g(x)<+\infty}} f(x) + \frac{\lambda}{2} g(x) \\ &= \inf_{\substack{r>0 \\ s\in S_g}} r f(s) + \frac{\lambda}{2} r^2 \\ &= \inf_{s\in S_g} \inf_{r>0} r f(s) + \frac{\lambda}{2} r^2 \\ &= \inf_{s\in S_g} \frac{-1}{2\lambda} \big(\min(f(s), 0)\big)^2 \\ &= -\frac{1}{2\lambda} \sup_{s\in S_g} (\min(f(s), 0))^2 . \end{aligned} \tag{I.2.35}$$

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

Furthermore, it is of course trivial that adding the regularisation in $\|\mathcal{A}p_f\|^2$ cannot increase the value of $\|\mathcal{A}p_f\|$ at the minimiser (if it exists) – in practice, the regularisation causes a significant decrease.

Remark I.29. In this section, we have focused on the Fenchel duality formulation of the problem of existence of a point in $K\mathcal{E}\cap\mathcal{Y}$. Notice that other algorithms could be considered to tackle this problem, such as orthogonal projection based algorithms or minimisations of appropriate functionals under constraints – for example, one could implement a projected gradient descent or a Frank-Wolfe algorithm on the minimisation problem

$$\inf_{u\in\mathcal{E}} \frac{1}{2} \|Ku - \Pi_{\mathcal{Y}}(Ku)\|^2. \tag{I.2.36}$$

However, it so happens that the Chambolle-Pock algorithm [CP11] is very well adapted to our control theory setting, for it has good convergence rates, immediately provides the appropriate dual certificate if $K\mathcal{E} \cap \mathcal{Y} = \emptyset$ and allows for intuitive regularisations. All in all, it is ideal for the multi-faceted computer-assisted proofs we develop.

I.2.5 Convex sets, polars and polytopes

We shall not prove the results in this section: we refer to classical textbooks such as [Roc70], [BC11] or [Sch13] for proofs. In this section only, we will consider $X = \mathbb{R}^n$. The results we introduce shall only be useful for Chapter IV.

I.2.5.a Convex sets and duality

Definition I.30. Consider a set $C \subset X$. C is said to be convex if

$$\forall\,(x,y) \in C,\ \forall\,\lambda \in [0,1], \qquad \lambda x + (1-\lambda)y \in C. \tag{I.2.37}$$

An *extreme point* of C is an element $z \in C$ that cannot be written as the convex combination $\lambda x + (1-\lambda)y$ for any $x, y \in C$ and $\lambda \in (0,1)$.

Recall as well the definition of the convex hull:

Definition I.31. Let $A \subset \mathbb{R}^n$. The convex hull of A is denoted $\mathrm{conv}(A)$ and defined as

$$\mathrm{conv}(A) = \left\{ x \in \mathbb{R}^n, \quad \exists\, m \in \mathbb{N}^*, \exists\,(x_i)_i \in A^m, \exists\,(\lambda_i)_i \in (\mathbb{R}_+)^n, \sum_{i=1}^m \lambda_i = 1, \sum_{i=1}^m \lambda_i x_i = x \right\}.$$

Proposition I.32. Let $A \subset \mathbb{R}^n$. The convex hull $\mathrm{conv}(A) \subset \mathbb{R}^n$ is always convex, and closed if A has finitely many elements. Furthermore,

$$\mathrm{conv}(A) = \bigcap_{\substack{A \subset C \subset \mathbb{R}^n \\ C \text{ is convex}}} C. \tag{I.2.38}$$

Proposition I.33. Let $C \subset \mathbb{R}^n$ a nonempty, closed, bounded convex set. Then it is equal to the convex hull of its extreme points.

Recall Theorem I.19, which hints that the study of convex sets is equivalent to the study of its support function (for more details about this equivalence, see [BC11, Proposition 14.11]). We now define the *polar sets* of convex sets, and study the duality between a convex and its polar.

Definition I.34. Let C be a nonempty set. Its polar C° is defined as the 1-sublevel set of the support function of C, that is:

$$C^\circ = \{x \in \mathbb{R}^n,\ \sigma_C(x) \leq 1\}. \tag{I.2.39}$$

The polar of its polar is denoted $C^{\circ\circ}$.

Let us list a few interesting properties of polar sets:

Theorem I.35. Let $C \subset \mathbb{R}^n$. Then C° is convex, closed and $0 \in C^\circ$. Moreover,

$$C^{\circ\circ} = \overline{\mathrm{conv}(C \cup \{0\})}. \tag{I.2.40}$$

It follows that if C is convex, closed and $0 \in C$, then $C^{\circ\circ} = C$. Finally, we have the equivalence:

$$0 \in \mathrm{Int}(C) \quad \Longleftrightarrow \quad C^\circ \text{ is bounded.} \tag{I.2.41}$$

Proposition I.36. Let $C_1, C_2 \subset \mathbb{R}^n$. If $C_1 \subset C_2$, then $C_2^\circ \subset C_1^\circ$. If both C_1 and C_2 are convex, closed and contain 0, we furthermore have that

$$(C_1 \cap C_2)^\circ = \mathrm{conv}(C_1^\circ \cup C_2^\circ) \quad \text{and} \quad (C_1^\circ \cup C_2^\circ)^\circ = C_1 \cap C_2. \tag{I.2.42}$$

Proposition I.37. Let $C \subset \mathbb{R}^n$ be a compact convex such that $0 \in \mathrm{Int}(C)$. Then the same goes for C°, and

$$\forall\, p \in (\mathbb{R}^n)^*, \quad \sigma_C(p) > 0 \text{ and } \sigma_{C^\circ}(p) > 0. \tag{I.2.43}$$

Furthermore,

$$\forall\, p \in (\mathbb{R}^n)^*, \exists\, q \in (\mathbb{R}^n)^*, \quad \left\langle \frac{p}{\sigma_C(p)}, \frac{q}{\sigma_{C^\circ}(q)} \right\rangle = 1 \tag{I.2.44}$$

In addition, $\frac{p}{\sigma_C(p)} \in \partial C^\circ$ and $\frac{q}{\sigma_{C^\circ}(q)} \in \partial C$.

I.2.5.b Convex polytopes

Among convex sets, we shall focus on convex polytopes, which are defined as follows:

Definition I.38. A convex polytope $P \subset \mathbb{R}^n$ is a closed convex set with a finite number of extreme points – its vertices.

Since we shall only deal with convex polytopes, we shall simply refer to them as polytopes. Polytopes have two useful characterisations:

Proposition I.39. A polytope can be characterised as the intersection of a finite number of half-spaces (its $\mathcal{H}$-representation): given $(p_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m$ its outer directions and $(b_j) \in \mathbb{R}^m$:

$$P = \bigcap_{j\in\{1,\dots,m\}} \{x \in \mathbb{R}^n, \langle p_j, x\rangle \leq b_j\}.$$

If furthermore P is bounded, one can equivalently characterise it using its vertices (its $\mathcal{V}$-representation): given its vertices $(y_i)_{i\in\{1,\dots,k\}} \in (\mathbb{R}^n)^k$, P is their convex hull

$$P = \mathrm{conv}((y_i)_i)$$

Note that even though those two representations of bounded polytopes are equivalent, the passage from one to another is a complicated task, referred as the vertex enumeration problem (from $\mathcal{H}$-representation to $\mathcal{V}$-representation) and facet enumeration problem (from $\mathcal{V}$-representation to $\mathcal{H}$-representation).

Remark as well that the $\mathcal{H}$-representation needs not be unique, but there exists a minimum number d of directions needed to define any polytope, and any $\mathcal{H}$-representation having exactly d directions is equal up to a positive multiplicative coefficient for each couple (p_j, b_j) and a permutation on their order.

We shall use the following non-standard vocabulary:

Definition I.40. Let $P \subset \mathbb{R}^n$ be a bounded polytope, with vertices $y = (y_i)_{i\in\{1,\dots,k\}} \in (\mathbb{R}^n)^k$ and $\mathcal{H}$-representation: given $(p_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m$ and $(b_j) \in \mathbb{R}^m$. To each direction (p_j) is associated a *facet* F_j of the polytope:

$$F_j = P \cap \{x \in \mathbb{R}^n,\ \langle p_j, x\rangle = b_j\} = \mathrm{conv}((v_j^i)_i), \tag{I.2.45}$$

where $(v_i^j)_i \in y$ are the vertices bordering the facet.

Note that a facet might be contained in a d-dimensional affine subspace, with $0 \leq d \leq n-1$, and so this definition does not align with other definitions in the literature where a facet is necessarily of dimension $n-1$. We shall now introduce a few tools essential to the study of polytopes: the support functions and the polar of a polytope.

Proposition I.41. Let $P \subset \mathbb{R}^n$ be a polytope defined by its $\mathcal{V}$-representation $(y_i)_{i\in\{1,\dots,k\}}$. Then

$$\forall\, x \in \mathbb{R}^n, \quad \sigma_P(x) = \max_{i\in\{1,\dots,k\}} \langle y_i, x\rangle. \tag{I.2.46}$$

Let us give a few properties of polytopes:

Proposition I.42. Let P_1, P_2 two polytopes. Then

- $P_1 \cap P_2$ is a polytope, and a (not necessarily optimal) $\mathcal{H}$-representation of $P_1 \cap P_2$ is the concatenation of their $\mathcal{H}$-representations
- if $P_1 \cup P_2$ is a convex, then it is a polytope, and its set of vertices is included in the union of the vertices of P_1 and P_2.

For polytopes, we can characterise the polar sets more precisely:

Proposition I.43. Let P be a polytope and P° its polar.
Assume $(y_i)_{i\in\{1,\dots,k\}}$ are the vertices of P and $(p_j, b_j)_{j\in\{1,\dots,m\}}$ its $\mathcal{H}$-representation. Then

- P° is a polytope.
- if P is bounded, the $\mathcal{H}$-representation of P° is given by the $(y_i)_i$:

$$P^\circ = \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n, \quad \langle y_i, x\rangle \leq 1\}$$

- if $0 \in P$, then $P^{\circ\circ} = P$ and the vertices of P° are $\left\{\frac{p_j}{b_j} \text{ s.t. } b_j > 0, j \in \{1,\dots,m\}\right\}$.
- if $0 \in \operatorname{Int} P$, then $P^{\circ\circ} = P$, P° is bounded and its $\mathcal{V}$-representation is $(\frac{p_j}{b_j})_j$.

For an illustration in the case where P is bounded and contains 0 in its interior, see Figure I.1.

I.3 Approximation theory

In this section we shall focus on introducing some elements of approximation theory that will be needed in Chapter III. Let X and $V \subset X$ be two real or complex Hilbert spaces, with V densely and continuously embedded into X, equipped with inner products $\langle\cdot,\cdot\rangle_V$, $\langle\cdot,\cdot\rangle_X$ and associated norms $\|\cdot\|_V$ and $\|\cdot\|_X$. We also identify X with its dual X' to obtain the Gelfand' triplet $V \subset X \subset V'$.

We will first introduce without proofs fundamental results of approximation of V with a finite-dimensional subspace $V_h \subset V$. For proofs and more detailed results, we refer to standard textbooks such as [QSS06] or [BCL99]. We will then present a few choices of approximations spaces and give details about why and where they are useful in this thesis.

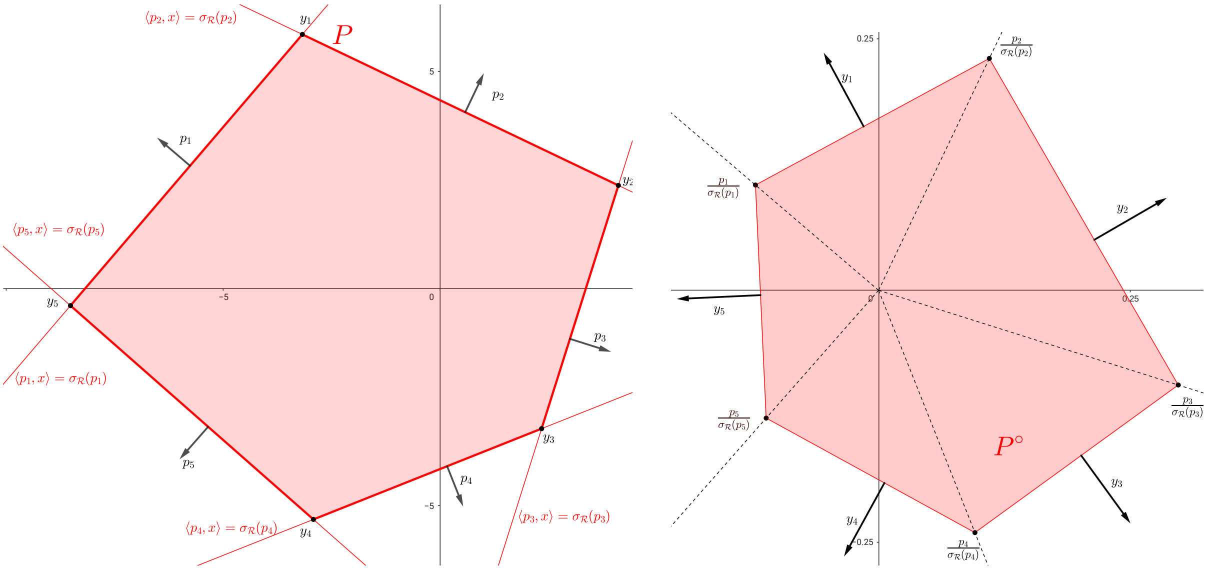

$\langle p_2, x\rangle = \sigma_{\mathcal{R}}(p_2)$
y_1
P
5
p_2
p_1
y_2
$\langle p_5, x\rangle = \sigma_{\mathcal{R}}(p_5)$
y_5
−5
0
p_3
$\langle p_1, x\rangle = \sigma_{\mathcal{R}}(p_1)$
p_5
y_3
p_4
−5
$\langle p_4, x\rangle = \sigma_{\mathcal{R}}(p_4)$
y_4
$\langle p_3, x\rangle = \sigma_{\mathcal{R}}(p_3)$
0.25
$\frac{p_2}{\sigma_{\mathcal{R}}(p_2)}$
y_1
$\frac{p_1}{\sigma_{\mathcal{R}}(p_1)}$
y_2
y_5
0
0.25
$\frac{p_3}{\sigma_{\mathcal{R}}(p_3)}$
$\frac{p_5}{\sigma_{\mathcal{R}}(p_5)}$
P°
y_3
y_4
−0.25
$\frac{p_4}{\sigma_{\mathcal{R}}(p_4)}$

Figure I.1: Example of a polytope (left panel) and its polar (right panel).

I.3.1 Fundamental results

Consider a bilinear continuous form $a(\cdot,\cdot)$ on V. Let $a_1 > 0$ be its continuity constant and furthermore assume them to be coercive with constant $0 < a_0 \leq a_1$:

$$\begin{cases} |a(u,v)| \leq a_1 \|u\|_V \|v\|_V & \forall\, u, v \in V \quad \text{(continuity)} \\ \operatorname{Re} a(v,v) \geq a_0 \|v\|_V^2 & \forall\, v \in V \quad \text{(coercivity).} \end{cases} \tag{I.3.1}$$

Theorem I.44 (Lax-Milgram)**.** Let $f \in V'$. Under assumptions I.3.1, there exists a unique $u \in V$ satisfying

$$a(u,v) = f(v) \qquad \forall\, v \in V. \tag{I.3.2}$$

Using Riesz's representation theorem, we can define of a bijective linear map $\mathcal{A} \in \mathcal{L}(V, V')$ through

$$a(u,v) = \langle \mathcal{A}u, v\rangle_V \qquad \forall\, u, v \in V, \tag{I.3.3}$$

that is furthermore continuous and coercive in V, in the sense that

$$\begin{cases} |\langle \mathcal{A}u, v\rangle| \leq a_1 \|u\|_V \|v\|_V & \forall\, u, v \in V \quad \text{(continuity)} \\ \operatorname{Re} \langle \mathcal{A}v, v\rangle \geq a_0 \|v\|_V^2 & \forall\, v \in V \quad \text{(coercivity).} \end{cases} \tag{I.3.4}$$

Note that this operator $\mathcal{A}$ of this section is in essence the opposite of the one used in section I.1. From now on we also consider a finite-dimensional subspace $V_h \subset V$ that aims at approximating V when $h \to 0$.

Lemma I.45 (Céa)**.** Let $u \in V$ and $f = \mathcal{A}u \in V'$, that is, they satisfy

$$a(u,v) = \langle f, v\rangle \qquad \forall\, v \in V, \tag{I.3.5}$$

and let furthermore $u_h \in V_h$ solve

$$a(u_h, v) = \langle f, v_h\rangle \qquad \forall\, v_h \in V_h. \tag{I.3.6}$$

Under assumptions I.3.1, we have that

$$\|u - u_h\|_V \leq \frac{a_1}{a_0} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{I.3.7}$$

If $\mathcal{A}$ (or a) is symmetric (or sesquilinear in the complex case), the following version of Céa's lemma holds:

Lemma I.46 (Céa)**.** Under the same notations and hypotheses that Lemma I.45, and if furthermore $a(\cdot,\cdot)$ is symmetric, then the following sharper estimate holds

$$\|u-u_h\|_V \leq \sqrt{\frac{a_1}{a_0}} \inf_{v_h\in V_h} \|u-v_h\|_V. \tag{I.3.8}$$

In particular, should one define a bijective operator $\mathcal{A}_h : V_h \to V_h$ defined by

$$\forall\, u_h, v_h \in V_h, \qquad \mathcal{A}_h u_h \in V_h \quad \text{and} \quad \langle \mathcal{A}_h u_h, v_h\rangle = \langle \mathcal{A} u_h, v_h\rangle, \tag{I.3.9}$$

and also define $P_h : X \to V_h$ the orthogonal projection on V_h with respect to the inner product of V – this is well defined because $X \subset V'$ and the inner product on V and duality product on $V' \times V$ identify. In that case, for all

$$u \in \mathcal{D}(\mathcal{A}) := \left\{ v\in V, \mathcal{A}v \in X \right\}, \tag{I.3.10}$$

$u_h = \mathcal{A}_h^{-1} P_h \mathcal{A} u \in V_h$ satisfies (I.3.6) and therefore is a suitable candidate for Cea's lemma, and is a quasi-optimal approximation of u in V_h. To make this result more precise, we shall need another assumption in the form of the following assumption on V_h: there exists $C_0 > 0$ such that

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

A slightly stronger form – yet more intuitive – of ($\mathcal{V}_1$) is the following

$$\forall\, u\in \mathcal{D}(A), \quad \begin{cases} \inf_{v_h\in V_h} \|u-v_h\|_V \leq C_0'\, h \|Au\|_X \\ \inf_{v_h\in V_h} \|u-v_h\|_V \leq C_0''\, h \|A^*u\|_X, \end{cases} \tag{I.3.11}$$

which is to be understood as V_h approximating elements of regularity $\mathcal{D}(\mathcal{A}) \cup \mathcal{D}(\mathcal{A}^*)$ in a $\mathcal{O}(h)$ manner in the V-norm. These kinds of assumptions are very standard and can be found abundantly in the literature (e.g. [QSS06]), and have intuitives (though fairly tedious) proofs in the considered examples (they follow from I.51 for $\mathbb{P}_1$ finite elements, and I.58 for cubic splines). Notice that when the value of the constants matter, (I.3.11) might provide smaller constants.

The following lemma proves that a consequence of ($\mathcal{V}_1$) is the approximation in a $\mathcal{O}(h^2)$ manner in the weaker X-norm for $u \in \mathcal{D}(\mathcal{A})$. In a similar manner, one could prove the same lemma for $u \in \mathcal{D}(\mathcal{A}^*)$.

Lemma I.47. Under ($\mathcal{V}_1$) and the preceding hypotheses on V_h and $\mathcal{A}_h$, we have

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

where

$$C_1 = \frac{a_1^2 C_0^2}{a_0} \quad \Big(\text{resp. } C_1 = \frac{a_1^{3/2} C_0^2}{4\sqrt{a_0}} \text{ if } (\mathcal{A}, \mathcal{D}(\mathcal{A})) \text{ is self-adjoint}\Big). \tag{I.3.12}$$

Proof. We shall only prove the lemma in the first case. If $\mathcal{A}$ is self-adjoint, the adjusted constant derives from the exact same method but using Céa's Lemma I.46 which yields $C_c = \sqrt{\frac{a_1}{a_0}}$, and noticing that ($\mathcal{V}_1$) simplifies into

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

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

$$\forall\, (v, w_h) \in X \times V_h, \quad P_h v \in V_h \quad \text{and} \quad \langle P_h v, w_h\rangle = \langle v, w_h\rangle. \tag{I.3.13}$$

Let $f \in X$. Applying Céa's Lemma I.45 to the maps

$$a : \begin{cases} V \times V \to \mathbb{C} \\ (v, w) \mapsto \langle \mathcal{A}v, w\rangle \end{cases} \qquad L : \begin{cases} V \to \mathbb{C} \\ v \mapsto \langle v, f\rangle \end{cases} \tag{I.3.14}$$

we get

$$\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\|_V \leq C_c \inf_{v_h \in V_h} \|\mathcal{A}^{-1}f - v_h\|, \tag{I.3.15}$$

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

$$\begin{aligned} \langle \mathcal{A}(\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf), v_h\rangle &= \langle f, v_h\rangle - \langle \mathcal{A}\mathcal{A}_h^{-1}P_hf, v_h\rangle \\ &= \langle f, v_h\rangle - \langle \mathcal{A}_h\mathcal{A}_h^{-1}P_hf, v_h\rangle \text{ since } \mathcal{A}_h^{-1}P_hf \in V_h \\ &= \langle f, v_h\rangle - \langle P_hf, v_h\rangle \\ &= 0. \end{aligned} \tag{I.3.16}$$

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

$$\begin{aligned} \langle g, \mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\rangle &= \langle \mathcal{A}^*(\mathcal{A}^*)^{-1}g, \mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\rangle \\ &= \langle (\mathcal{A}^*)^{-1}g - v_h, \mathcal{A}(\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf)\rangle \quad \forall v_h \in V_h \text{ using (I.3.16)} \\ &\leq a_1\|(\mathcal{A}^*)^{-1}g - v_h\|_V\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\|_V \quad \forall v_h \in V_h \\ &\leq a_1\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\|_V \inf_{v_h \in V_h} \|(\mathcal{A}^*)^{-1}g - v_h\|_V. \end{aligned} \tag{I.3.17}$$

Hence

$$\begin{aligned} \|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\| &= \sup_{\substack{g \in X \\ \|g\|=1}} \langle g, \mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\rangle \\ &\leq a_1\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\|_V \sup_{\|g\|=1} \inf_{v_h \in V_h} \|(\mathcal{A}^*)^{-1}g - v_h\|_V. \end{aligned} \tag{I.3.18}$$

Using the bounds ($\mathcal{V}_1$) and (I.3.15), we obtain

$$\begin{aligned} \|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\| &\leq a_1\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\|_V \sup_{\|g\|=1} C_0h\|g\| \text{ using } (\mathcal{V}_1) \\ &= a_1C_0h\|\mathcal{A}^{-1}f - \mathcal{A}_h^{-1}P_hf\|_V \\ &\leq a_1C_0C_ch\sqrt{a_0}h \inf_{v_h \in V_h} \|\mathcal{A}^{-1}f - v_h\|_V \text{ using (I.3.15)} \\ &\leq a_1C_0^2C_ch^2\|f\| \text{ using } (\mathcal{V}_1). \end{aligned} \tag{I.3.19}$$

The result is proved, with $C_1 = a_1C_0^2C_c = \frac{a_1^2C_0^2}{a_0}$. □

In order to implement the operator $\mathcal{A}$ on V, one only needs to define the discretised operator $\mathcal{A}_h$ satisfying (I.3.9). This is the objective of the following standard proposition (for details, see e.g. [Tho07, Page 7]).

Proposition I.48. Define $(\varphi_i)_{i\in\{1,\ldots,\dim V_h\}}$ a basis of V_h, and

$$I_h : \begin{cases} \mathbb{R}^{\dim V_h} & \to V_h \\ z & \mapsto \sum_{i=1}^{N-1} z_i\varphi_i. \end{cases} \tag{I.3.20}$$

I_h is a bijection from $\mathbb{R}^{\dim V_h}$ to V_h. Furthermore, define the two $(N-1)\times(N-1)$ matrices

$$M = (\langle \varphi_i, \varphi_j\rangle_X)_{(i,j)\in\{1,\ldots,N-1\}} \quad \text{and} \quad K = (\langle A\varphi_i, \varphi_j\rangle_X)_{(i,j)\in\{1,\ldots,N-1\}}. \tag{I.3.21}$$

Henceforward, M will be called the *mass matrix* and K the *stiffness matrix*. The operator defined as

$$\mathcal{A}_h : \begin{cases} V_h & \to V_h \\ u_h & \mapsto I_h(M^{-1}KI_h^{-1}u_h) \end{cases} \tag{I.3.22}$$

is a discretisation of $\mathcal{A}$ on V_h, in the sense that

$$\forall\, u_h, v_h \in V_h, \qquad \mathcal{A}_h u_h \in V_h \quad \text{and} \quad \langle \mathcal{A}_h u_h, v_h\rangle = \langle \mathcal{A}u_h, v_h\rangle. \tag{I.3.23}$$

In the following sections, we shall focus on the case of the 1D heat equation, which is the most common and useful scenario studied in Section III.3.3.

I.3.2 Approximation of the 1D heat equation

Let $X = L^2(0,1)$, $V = H_0^1(0,1) \subset X$. In this section, the bilinear form considered is defined as

$$a(\cdot,\cdot) : \begin{cases} V \times V & \to \mathbb{R} \\ (u,v) & \mapsto a(u,v) = \langle u, v\rangle_V = \langle u', v'\rangle_X. \end{cases} \tag{I.3.24}$$

As has been seen before, this bilinear form can be associated to a unique linear operator $\mathcal{A} : V \to V'$ such that

$$\forall\,(u,v) \in V^2, \qquad a(u,v) = \langle \mathcal{A}u, v\rangle. \tag{I.3.25}$$

On the subspace $H_0^1(0,1) \cap H^2(0,1) =: \mathcal{D}(\mathcal{A})$, this definition coincide with the opposite of the Dirichlet laplacian with zero Dirichlet boundary conditions, that is, in our context:

$$\forall\, u \in H_0^1(0,1) \cap H^2(0,1), \qquad \mathcal{A}u = -u''. \tag{I.3.26}$$

In particular, $\mathcal{A}$ is self-adjoint.

Proposition I.49. $\mathcal{A}$ is continuous and coercive with constants $a_0 = a_1 = 1$.

Proof. Let $u, v \in \mathcal{D}(\mathcal{A}) = H_0^1(0,1) \cap H^2(0,1)$. Then

$$|\langle \mathcal{A}u, v\rangle| = |\langle \nabla u, \nabla v\rangle| \leq \|u\|_V \|v\|_V, \tag{I.3.27}$$

and

$$\operatorname{Re}\langle \mathcal{A}v, v\rangle = \operatorname{Re}\langle \nabla v, \nabla v\rangle = \|v\|_V^2. \tag{I.3.28}$$

□

As we have seen, to compute approximations of the solutions to the heat equation, one must choose the discretisation space V_h on which to discretise $\mathcal{A}$ into $\mathcal{A}_h$. The choice of V_h must be made while keeping in mind the specific setting and aim of the discretisation. In our case, the setting of II, that is computer-assisted proofs, require a careful treatment of errors, including:

- a "precise" approximation, meaning with small overall discretisation errors, both in rates of convergence in time and space and in small constants associated to thoses rates
- a "well-conditioned" setting, that will keep round-off errors small even when considering very fine discretisations
- the computation of $\|\mathcal{A}p_f\|_X$ for a given p_f, obtained first through a minimisation process and potentially through interpolation.

Two general options of discretisation emerge from these criteria : the first is the "simple" one, consisting in discretising as simply as possible, for the simpler the space the smaller the round-off errors will be. The primary example of such a space would be the $\mathbb{P}_1$ finite elements, which have two apparent drawbacks: their small order of convergence ($\mathcal{O}(h)$ in V-norm) and the fact that $V_h \not\subset \mathcal{D}(\mathcal{A})$, which implies the need of interpolation in order to compute $\|\mathcal{A}p_f\|_X$.

The other option could be called the "convenient" one: contrarily to the first, it is regular enough so that $V_h \subset \mathcal{D}(\mathcal{A})$, and its higher complexity allows for a higher order of convergence. The primary example of this option would be cubic splines, which could converge in $\mathcal{O}(h^3)$. However, it has two main drawbacks: first, its complexity on one hand increases the dimension of the subspace and the quantity of numerical computations (and thus, of round-off errors), and on the other hand the order of convergence $\mathcal{O}(h^3)$ requires a higher regularity of the minimiser p_f, namely $p_f \in \mathcal{D}(\mathcal{A}^2)$. Since cubic splines are not that regular, they would themselves require interpolation into, for example, heptic splines that are in $\mathcal{D}(\mathcal{A})$. But then the same remark can apply: why not use heptic splines directly to obtain even better rates of convergence?

In an other direction, one could also consider requiring less regularity, for example $p_f \in \mathcal{D}(\mathcal{A}^{1/2})$, at the cost of lower rates of convergence. However, the rates thus obtained (essentially $\mathcal{O}(h)$ in X-norm) would be prohibitive considering the extra-small error bounds we need for computer-assisted proofs. Therefore, in the following sections we will mainly introduce $\mathbb{P}_1$ finite elements, which were the main method to obtain computer-assisted proofs in this thesis. We will also introduce the cubic and heptic splines and their surrounding context – cubic splines will be used in Chapter III for interpolation purposes, and heptic splines could be used to interpolate cubic splines.

I.3.2.a Finite elements

In this section we will focus on $\mathbb{P}_1$ finite elements in the one-dimensional case. These are the simplest form of finite elements, and we shall here give a few well known lemmas, as well as provide the proof with explicit constants that we require for the proofs of Chapter II. For a more extensive study, we refer to standard textbooks, such as [QSS06; Tho07].

Let $N \in \mathbb{N}^*$, $h = \frac{1}{N}$ and V_h the space of $\mathbb{P}_1$ finite elements on the grid $\{x_i = ih, i \in \{0, \dots, N\}\}$:

$$V_h = \{f \in V, \quad f(0) = f(1) = 0 \text{ and } \forall\, i \in \{0, \dots, N-1\}, f \text{ is affine on } [x_i, x_{i+1}]\}. \qquad \text{(I.3.29)}$$

Notice that one could just as well use a non-uniform mesh, and define $h = \max_i |x_{i+1} - x_i|$. Denoting

$$\forall\, i \in \{1, \dots, N-1\}, \qquad \varphi_i : x \mapsto \begin{cases} 0 & \text{if } x \le x_{i-1} \\ \frac{1}{h}(x - x_{i-1}) & \text{if } x_{i-1} \le x \le x_i \\ \frac{1}{h}(x_{i+1} - x) & \text{if } x_i \le x \le x_{i+1} \\ 0 & \text{if } x_{i+1} \le x, \end{cases} \qquad \text{(I.3.30)}$$

then $(\varphi_i)_{i \in \{1,\dots,N-1\}}$ is a basis of V_h (see Figure I.2).

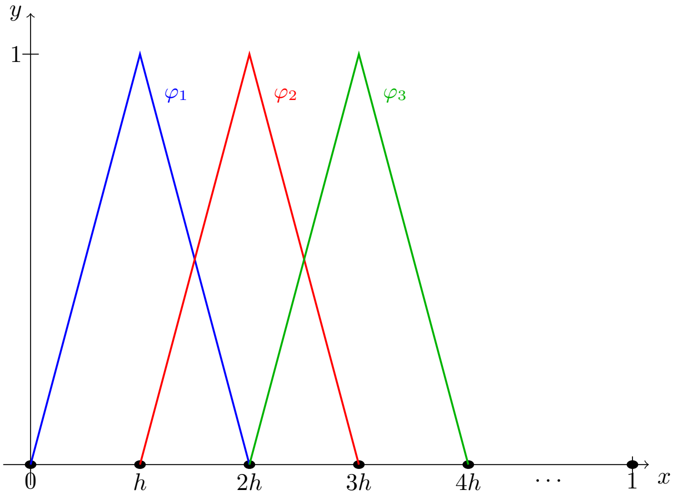


Figure I.2: P1 finite element basis functions.

Proposition I.50. The mass matrix has coefficients:

$$\forall i,j \in \{1,\dots,N-1\}, \qquad M_{i,j} = \langle \varphi_i, \varphi_j \rangle_X = \begin{cases} \frac{2h}{3} & \text{if } i=j \\ \frac{h}{6} & \text{if } |i-j|=1 \\ 0 & \text{if } |i-j|\geq 2, \end{cases} \tag{I.3.31}$$

and the stiffness matrix has coefficients:

$$\forall i,j \in \{1,\dots,N-1\}, \qquad K_{i,j} = \langle A\varphi_i, \varphi_j \rangle_X = \begin{cases} \frac{2}{h} & \text{if } i=j \\ -\frac{1}{h} & \text{if } |i-j|=1 \\ 0 & \text{if } |i-j|\geq 2. \end{cases} \tag{I.3.32}$$

We shall now prove a standard approximation result using $\mathbb{P}_1$ finite elements, which in turn proves the assumption ($\mathcal{V}_1$) for V_h.

Proposition I.51. Let $N \geq 1$, $h = \frac{1}{N}$ and $x_i = ih$ for $i \in \{0,\dots,N\}$. We have for all $f \in H^2(0,1) \cap H^1_0(0,1)$,

$$\left\| f - \sum_{i=1}^{N-1} f(x_i)\,\varphi_i \right\|_V \leq \frac{h}{\sqrt{2}} \|f''\|_X, \tag{I.3.33}$$

and

$$\left\| f - \sum_{i=1}^{N-1} f(x_i)\varphi_i \right\|_X \leq \frac{h^2}{2\sqrt{2}} \|f''\|_X. \tag{I.3.34}$$

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

$$\begin{aligned} \int_{x_i}^{x_{i+1}} |f'(x) - p'(x)|^2 \,\mathrm{d}x &= \int_{x_i}^{x_{i+1}} |f'(x) - f'(\theta_i)|^2 \,\mathrm{d}x = \int_{x_i}^{x_{i+1}} \left| \int_{\theta_i}^{x} f''(t)\,\mathrm{d}t \right|^2 \mathrm{d}x && \text{(I.3.35)} \\ &\leq \int_{x_i}^{x_{i+1}} |\theta_i - x| \int_{\theta_i}^{x} |f''(t)|^2 \,\mathrm{d}t\,\mathrm{d}x && \text{(I.3.36)} \\ &\leq \frac{1}{2}\left((x_{i+1}-\theta_i)^2 + (\theta_i - x_i)^2\right) \int_{x_i}^{x_{i+1}} |f''(t)|^2 \,\mathrm{d}t && \text{(I.3.37)} \\ &\leq \frac{1}{2}(x_{i+1}-x_i)^2 \int_{x_i}^{x_{i+1}} |f''(t)|^2 \,\mathrm{d}t. && \text{(I.3.38)} \end{aligned}$$

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

$$\int_{x_i}^{x_{i+1}} |f(x) - p(x)|^2 \,\mathrm{d}x = \int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} |f(x) - p(x)|^2 \,\mathrm{d}x + \int_{\frac{1}{2}(x_i+x_{i+1})}^{x_{i+1}} |f(x) - p(x)|^2 \,\mathrm{d}x, \tag{I.3.39}$$

where

$$\int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} |f(x)-p(x)|^2 \,\mathrm{d}x = \int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} \left|\int_{x_i}^{x} (f'(t)-p'(t))\,\mathrm{d}t\right|^2 \mathrm{d}x \tag{I.3.40}$$

$$\leq \int_{x_i}^{\frac{1}{2}(x_i+x_{i+1})} |x-x_i| \int_{x_i}^{x} |f'(t)-p'(t)|^2 \,\mathrm{d}t\,\mathrm{d}x \tag{I.3.41}$$

$$\leq \frac{1}{8}(x_{i+1}-x_i)^2 \int_{x_i}^{x_{i+1}} |f'(t)-p'(t)|^2 \,\mathrm{d}t \tag{I.3.42}$$

$$\leq \frac{(x_{i+1}-x_i)^4}{16} \int_{x_i}^{x_{i+1}} |f''(t)|^2 \,\mathrm{d}t. \tag{I.3.43}$$

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

$$\int_{x_i}^{x_{i+1}} |f(x)-p(x)|^2 \,\mathrm{d}x \leq \frac{(x_{i+1}-x_i)^4}{8} \int_{x_i}^{x_{i+1}} |f''(t)|^2 \,\mathrm{d}t. \tag{I.3.44}$$

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

Since $\mathcal{A}$ is self-adjoint, using (I.3.33), one can easily see that $(\mathcal{V}_1)$ is satisfied with $C_0 = \sqrt{2}$.

As will be detailed in Chapter III, the main drawback of $\mathbb{P}_1$ finite elements lies in its limited regularity. In particular, $V_h \not\subset \mathcal{D}(\mathcal{A})$, which becomes problematic in Chapter III, as error bounds on the discretisation depend on $\|\mathcal{A}p_f\|$ (see Proposition III.9). This will require interpolation of these functions, and as will be seen in the next section, splines, especially cubic splines, are an efficient way to do it.

I.3.2.b Splines: the general case

In this section, we shall study an interpolation method based on polynomial by part functions: splines. We will not delve into the details of these methods, which have been extensively studied, and will refer to standard textbooks such as [De 01]. Given the same mesh as was used for finite elements, the n-spline space is the space of functions that are equal to polynomials of degree n on each interval of the mesh:

$$\$^h_n := \{\mathfrak{s} : [0,1] \to \mathbb{R}, \quad \forall\, i \in \{0,\dots,N-1\},\ \mathfrak{s}_{|[x_i,x_{i+1}]} \in \mathbb{R}_n([x_i,x_{i+1}])\}. \tag{I.3.45}$$

We shall now prove the following theorem, which turns out to be especially useful in justifying the use of splines in the context of interpolation $\mathbb{P}_1$ finite elements into the regular subspace $\mathcal{D}(\mathcal{A}) \subset V$: in Chapter III, the theoretical discretisation error bounds depend on the norm $\|\mathcal{A}p_f\|$.

Lemma I.52. Let $f \in H^m(0,1)$. Let $\mathfrak{s}_f$ be the unique $(2m-1)$-spline such that

$$\mathfrak{s}_f \in H^m(0,1) \quad \text{and} \quad \forall\,(i,j) \in \{0,\dots,N\} \times \{0, m-1\},\ \mathfrak{s}_f^{(j)}(x_i) = f^{(j)}(x_i). \tag{I.3.46}$$

Then

$$\|f^{(m)}\|^2_{L^2(0,1)} = \|\mathfrak{s}_f^{(m)}\|^2_{L^2(0,1)} + \|(f-\mathfrak{s}_f)^{(m)}\|^2. \tag{I.3.47}$$

Proof. We have

$$\begin{aligned}
\|f^{(m)}\|^2_{L^2(0,1)} &= \|f^{(m)}\|^2 \\
&= \|\mathfrak{s}_f^{(m)} + (f-\mathfrak{s}_f)^{(m)}\|^2 \\
&= \|\mathfrak{s}_f^{(m)}\|^2 + 2\langle \mathfrak{s}_f^{(m)}, (f-\mathfrak{s}_f)^{(m)}\rangle + \|(f-\mathfrak{s}_f)^{(m)}\|^2.
\end{aligned}$$

Moreover, $\mathfrak{s}_f$ and $f - \mathfrak{s}_f$ are both $\mathcal{C}^{m-1}$ on each interval (x_i, x_{i+1}). Hence, doing m integration by parts, we obtain

$$\begin{aligned}
\langle \mathfrak{s}_f^{(m)}, (f - \mathfrak{s}_f)^{(m)} \rangle &= \sum_{i=0}^{N_x} \int_{x_i}^{x_{i+1}} \mathfrak{s}_f^{(m)}(x)(f - \mathfrak{s}_f)^{(m)}(x)\,\mathrm{d}x \\
&= \sum_{i=0}^{N_x} \Bigg[(-1)^m \int_{x_i}^{x_{i+1}} \mathfrak{s}_f^{(2m)}(x)(f - \mathfrak{s}_f)(x)\,\mathrm{d}x \\
&\qquad + \sum_{j=0}^{n-1} \left[\mathfrak{s}_f^{(m+j)}(x)(f - \mathfrak{s}_f)^{(m-1-j)}(x) \right]_{x_i}^{x_{i+1}} \Bigg],
\end{aligned}$$

which yields 0 since on one hand $\mathfrak{s}_f$ is a $(2m-1)$spline, and hence its $2m$-derivative is zero, and on the other hand by definition of $\mathfrak{s}_f$ all the boundary terms cancel out. And finally

$$\|f^{(m)}\|^2_{L^2(0,1)} = \|\mathfrak{s}_f^{(m)}\|^2_{L^2(0,1)} + \|(f - \mathfrak{s}_f)^{(m)}\|^2. \tag{I.3.48}$$

□

Theorem I.53. Let $m \in \mathbb{N}^*$, let $q = (q_i^j)_{(i,j)\in\{0,\dots,N\}\times\{0,\dots,m-1\}}$, and given indices $\mathcal{I} \subset \{0,\dots,N\} \times \{0,\dots,m-1\}$. Then among the m-times differentiable functions f matching q in the following manner

$$\forall\,(i,j) \in \mathcal{I}, \quad f^{(j)}(x_i) = q_i^j, \tag{I.3.49}$$

$(2m-1)$-splines will provide the L^2-optimal option. More formally, we have that

$$\inf_{\substack{f \in H^m(0,1) \\ \forall\,(i,j)\in\mathcal{I},\, f^{(j)}(x_i) = q_i^j}} \|f^{(m)}\|_{L^2} = \inf_{\substack{\mathfrak{s} \in \$^h_{2m-1} \cap H^m(0,1) \\ \forall\,(i,j)\in\mathcal{I},\, \mathfrak{s}^{(j)}(x_i) = q_i^j}} \|\mathfrak{s}^{(m)}\|_{L^2}. \tag{I.3.50}$$

Furthermore all minimisers – if they exist – are $(2m-1)$-splines.

Proof. Firstly, it follows immediately from Lemma I.52 that the minimiser, if it exists, is a $(2m-1)$-spline: indeed, should $f \in H^m(0,1)$ satisfy the constraints, then define $\mathfrak{s}_f$ to be the $(2m-1)$ spline interpolating it. It follows that

$$\|\mathfrak{s}_f^{(m)}\|^2_{L^2(0,1)} \leq \|\mathfrak{s}_f^{(m)}\|^2_{L^2(0,1)} + \|(f - \mathfrak{s}_f)^{(m)}\|^2 = \|f^{(m)}\|^2_{L^2(0,1)}. \tag{I.3.51}$$

Notice furthermore that this inequality becomes an equality if and only if $f^{(m)} = \mathfrak{s}_f^{(m)}$ which implies that f, $\mathfrak{s}_f$ are all equal up to a $(m-1)$-spline. This $(m-1)$-spline is bound to be zero because $\mathfrak{s}_f$ (resp. its derivatives) agrees with f (resp. its derivatives) at all points $(x_i)_i$. □

In the following sections, we shall focus on cubic splines (3-splines) and heptic splines (7-splines) because of the advantageous optimal interpolation property provided by Theorem I.53 with respect to operators Δ and Δ^2. These unfortunately only have theoretical value: see Remark III.14.

I.3.2.c Cubic splines

Here we shall focus cubic splines $\$^h_3$, or more specifically on $\mathcal{S}^h_3 \subset \$^h_3$ the subspace of $\mathcal{C}^1$ cubic splines on the usual intervals $[ih, (i+1)h]$ for $i \in \{0,\dots,N\}$. While $\$^h_3$ was of dimension $4N-2$ (recall that two dimensions are removed by the zero boundary conditions), and its usual basis would consist of monomials of degrees $0, 1, 2$ or 3 on each interval, $\mathcal{S}^h_3$ is of dimension $2N$, and one of its basis is decomposed into $(\psi_i^0)_{i\in\{1,\dots,N-1\}}$ determining the values of the spline on the mesh, and $(\psi_i^1)_{i\in\{0,\dots,N\}}$ determining the values of their derivative on the mesh: see Figures I.3 and I.4. These basis functions are formally defined as:

$$\forall i \in \{1, \ldots, N-1\}, \qquad \psi_i^0 : x \mapsto \begin{cases} 0 & \text{if } x \leq x_{i-1} \\ -\frac{2}{h^3}(x - x_{i-1})^2(x - x_i - \frac{h}{2}) & \text{if } x_{i-1} \leq x \leq x_i \\ \frac{2}{h^3}(x - x_i + \frac{h}{2})(x - x_{i+1})^2 & \text{if } x_i \leq x \leq x_{i+1} \\ 0 & \text{if } x \geq x_{i+1}, \end{cases} \tag{I.3.52}$$

$$\forall i \in \{1, \ldots, N-1\}, \qquad \psi_i^1 : x \mapsto \begin{cases} 0 & \text{if } x \leq x_{i-1} \\ \frac{1}{h^3}(x - x_{i-1})^2(x - x_i) & \text{if } x_{i-1} \leq x \leq x_i \\ \frac{1}{h^3}(x - x_i)(x - x_{i+1})^2 & \text{if } x_i \leq x \leq x_{i+1} \\ 0 & \text{if } x \geq x_{i+1}, \end{cases} \tag{I.3.53}$$

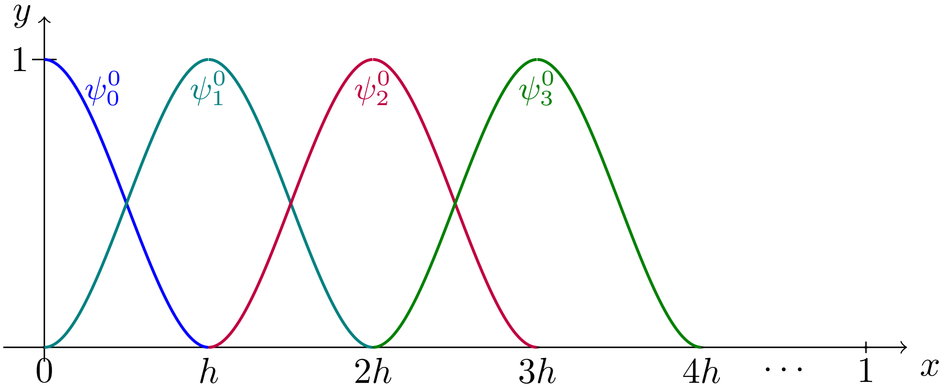


Figure I.3: Cubic splines value basis.

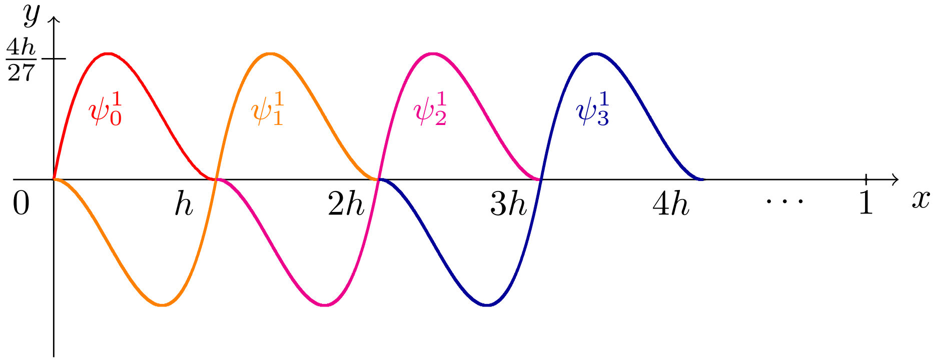


Figure I.4: Cubic splines derivative basis.

These functions form a basis of $\mathcal{S}_3^h$ because their total number is equal to the dimension of $\mathcal{S}_3^h$ and furthermore,

$$\forall (i,j) \in \{0, \ldots, N\}^2, \forall (k,l) \in \{0,1\}^2, \qquad (\psi_i^k)^{(l)}(x_j) = \begin{cases} 1 \text{ if } (i,k) = (j,l) \\ 0 \text{ otherwise.} \end{cases} \tag{I.3.54}$$

Even though they are not used in $\mathcal{S}_3^h$ – recall the zero boundary conditions – we also introduce the basis functions ψ_0^0 and ψ_N^0, defined as before.

Remark I.54. Note that many other bases are available for $\mathcal{C}^1$ cubic splines. The choice of a basis must be made while keeping in mind an important point: aside from its simplicity of expression, one must carefully consider the condition number of its mass and stiffness matrices: since every step shall be carried out using rigorous numerics, ill-conditioned matrices will significantly increase round-off error upper-bounding (see I.4 for details). This condition number is of course linked to the sparsity of the matrices, which in turn is linked to the locality of each basis function, i.e. the number of intervals on which it takes non-zero values. This basis is optimal in terms of the locality of each basis function, and therefore is a suitable candidate regarding the conditioning number.

In order to numerically deal with cubic splines, one should compute explicitely its mass and stiffness matrices.

Proposition I.55. The mass matrix associated to the basis $(\psi_i^j)_{i,j}$ can be computed using the following results:

$$\forall\,(i,j)\in\{0,\dots,N\},\qquad \langle\psi_i^0,\psi_j^0\rangle=\begin{cases}\frac{26}{35}h & \text{if } i=j\notin\{0,N\}\\ \frac{9}{70}h & \text{if } |i-j|=1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.55}$$

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle\psi_i^0,\psi_j^1\rangle=\begin{cases}0 & \text{if } i=j\notin\{0,N\}\\ \frac{-13}{420}h & \text{if } i=j-1\\ \frac{13}{420}h & \text{if } i=j+1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.56}$$

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle\psi_i^1,\psi_j^1\rangle=\begin{cases}\frac{2}{105}h & \text{if } i=j\notin\{0,N\}\\ \frac{-1}{140}h & \text{if } |i-j|=1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.57}$$

and finally $\langle\psi_0^0,\psi_0^0\rangle=\langle\psi_N^0,\psi_N^0\rangle=\frac{13}{35}h$, $\langle\psi_0^0,\psi_0^1\rangle=-\langle\psi_N^1,\psi_N^1\rangle=\frac{11}{210}h$, and $\langle\psi_0^1,\psi_0^1\rangle=\langle\psi_N^1,\psi_N^1\rangle=\frac{1}{105}h$.

Proposition I.56. The stiffness matrix associated to the basis $(\psi_i^j)_{i,j}$ can be computed using the following results:

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle\mathcal{A}\psi_i^0,\psi_j^0\rangle=\begin{cases}\frac{12}{5h} & \text{if } i=j\notin\{0,N\}\\ \frac{-6}{5h} & \text{if } |i-j|=1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.58}$$

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle\mathcal{A}\psi_i^0,\psi_j^1\rangle=\begin{cases}0 & \text{if } i=j\notin\{0,N\}\\ \frac{-1}{10h} & \text{if } i=j-1\\ \frac{1}{10h} & \text{if } i=j+1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.59}$$

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle\mathcal{A}\psi_i^1,\psi_j^1\rangle=\begin{cases}\frac{4}{15h} & \text{if } i=j\notin\{0,N\}\\ \frac{-1}{30h} & \text{if } |i-j|=1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.60}$$

and finally $\langle\mathcal{A}\psi_0^0,\psi_0^0\rangle=\langle\mathcal{A}\psi_N^0,\psi_N^0\rangle=\frac{6}{5h}$, $\langle\mathcal{A}\psi_0^0,\psi_0^1\rangle=\langle\mathcal{A}\psi_N^1,\psi_N^1\rangle=\frac{1}{10h}$, and $\langle\mathcal{A}\psi_0^1,\psi_0^1\rangle=\langle\mathcal{A}\psi_N^1,\psi_N^1\rangle=\frac{2}{15h}$.

Finally, as mentioned before, we shall need a closed formula to compute $\|\mathcal{A}p_f\|$ for $p_f\in\mathcal{S}_3^h$. The following proposition allows just that:

Proposition I.57. We have

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle \mathcal{A}\psi_i^0,\mathcal{A}\psi_j^0\rangle=\begin{cases}\frac{24}{h^3} & \text{if } i=j\\ \frac{-12}{h^3} & \text{if } |i-j|=1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.61}$$

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle \mathcal{A}\psi_i^0,\mathcal{A}\psi_j^1\rangle=\begin{cases}0 & \text{if } i=j\\ \frac{-6}{h^3} & \text{if } i=j-1\\ \frac{6}{h^3} & \text{if } i=j+1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.62}$$

$$\forall\,(i,j)\in\{1,\dots,N-1\},\qquad \langle \mathcal{A}\psi_i^1,\mathcal{A}\psi_j^1\rangle=\begin{cases}\frac{8}{h^3} & \text{if } i=j\\ \frac{2}{h^3} & \text{if } |i-j|=1\\ 0 & \text{if } |i-j|>1,\end{cases}\tag{I.3.63}$$

and finally $\langle \mathcal{A}\psi_0^0,\mathcal{A}\psi_0^0\rangle=\langle \mathcal{A}\psi_N^0,\mathcal{A}\psi_N^0\rangle=\frac{12}{h^3}$, $\langle \mathcal{A}\psi_0^0,\mathcal{A}\psi_0^1\rangle=-\langle \mathcal{A}\psi_N^1,\mathcal{A}\psi_N^1\rangle=\frac{6}{h^3}$, and $\langle \mathcal{A}\psi_0^1,\mathcal{A}\psi_0^1\rangle=\langle \mathcal{A}\psi_N^1,\mathcal{A}\psi_N^1\rangle=\frac{4}{h^3}$.

Approximation of V. These results allow the manipulation of $\mathcal{A}$ on $\mathcal{S}_3^h\subset V$, through a discretised operator, exactly as for the $\mathbb{P}_1$ finite elements. It is then natural to ask ourselves how the approximation properties of cubic splines compare to those of $\mathbb{P}_1$ finite elements. This is the objective of the following proposition. This proposition is done without assuming any boundary conditions, which is equivalent to adding ψ_0^0 and ψ_N^0 to the previous basis.

Proposition I.58. Let $N\geq 1$, $h=\frac{1}{N}$ and $x_i=ih$ for $i\in\{0,\dots,N\}$. We have for all $f\in H^2(0,1)$,

$$\left\| f-\sum_{i=0}^{N} f(x_i)\,\psi_i^0-\sum_{i=0}^{N} f'(x_i)\,\psi_i^1\right\|_V\leq \frac{h}{2}\|f''\|_X,\tag{I.3.64}$$

and

$$\left\| f-\sum_{i=0}^{N} f(x_i)\,\psi_i^0-\sum_{i=0}^{N} f'(x_i)\,\psi_i^1\right\|_X\leq \frac{h^2}{8}\|f''\|_X.\tag{I.3.65}$$

Furthermore, assuming further regularity $f\in H^4(0,1)$, we can get, for $k\in\{0,\dots,3\}$

$$\left\|\left(f-\sum_{i=0}^{N} f(x_i)\,\psi_i^0-\sum_{i=0}^{N} f'(x_i)\,\psi_i^1\right)^{(k)}\right\|_X\leq C_k h^{4-k}\|f^{(4)}\|_X,\tag{I.3.66}$$

where $C_0=\frac{3\sqrt{5}}{160}$, $C_1=\frac{\sqrt{7}}{8\sqrt{2}}$, $C_2=\frac{\sqrt{3}}{2}$ and $C_3=\frac{1}{2}$.

Proof. Assume first that $f\in H^2(0,1)$, and denote $\mathfrak{s}_f=\sum_{i=0}^N f(x_i)\,\psi_i^0+\sum_{i=0}^N f'(x_i)\,\psi_i^1$. We will only prove the various estimates on the interval $[0,\frac{h}{2}]$: the symmetry of hypotheses allows the same estimates on $[\frac{h}{2},h]$, and then on each interval $[x_i,x_{i+1}]$, and summing them and taking the square root will achieve the desired result.

$$\begin{aligned}\int_0^{\frac{h}{2}}(f'(x)-\mathfrak{s}_f'(x))^2\,\mathrm{d}x&=\int_0^{\frac{h}{2}}\left(\int_0^x f''(t)-\mathfrak{s}_f''(t)\,\mathrm{d}t\right)^2\mathrm{d}x\\&\leq\int_0^{\frac{h}{2}}x\int_0^x(f''(t)-\mathfrak{s}_f''(t))^2\,\mathrm{d}t\,\mathrm{d}x\quad\text{using Cauchy-Schwarz inequality}\\&\leq\int_0^{\frac{h}{2}}x\,\mathrm{d}x\int_0^h(f''(t)-\mathfrak{s}_f''(t))^2\,\mathrm{d}t.\end{aligned}$$

Recalling that using Lemma I.52, $\int_0^h (f''(t)-\mathfrak{s}_f''(t))^2 \,\mathrm{d}t = \int_0^h (f''(t))^2 \,\mathrm{d}t - \int_0^h (\mathfrak{s}_f''(t))^2 \,\mathrm{d}t \leq \int_0^h (f''(t))^2 \,\mathrm{d}t$, it follows that

$$\int_0^{\frac{h}{2}} (f'(x) - \mathfrak{s}_f'(x))^2 \,\mathrm{d}x \leq \frac{h^2}{8} \int_0^h (f''(t))^2 \,\mathrm{d}t, \tag{I.3.67}$$

which concludes the first estimate. For the second estimate, a similar method allows its easy proof:

$$\begin{aligned}
\int_0^{\frac{h}{2}} (f(x) - \mathfrak{s}_f(x))^2 \,\mathrm{d}x &= \int_0^{\frac{h}{2}} \left(\int_0^x f'(t) - \mathfrak{s}_f'(t) \,\mathrm{d}t \right)^2 \mathrm{d}x \\
&\leq \int_0^{\frac{h}{2}} x \int_0^x (f'(t) - \mathfrak{s}_f'(t))^2 \,\mathrm{d}t \,\mathrm{d}x \\
&= \int_0^{\frac{h}{2}} x \int_0^x \left(\int_0^t f''(u) - \mathfrak{s}_f''(u) \,\mathrm{d}u \right)^2 \mathrm{d}t \,\mathrm{d}x \\
&\leq \int_0^{\frac{h}{2}} x \int_0^x t \int_0^t (f''(u) - \mathfrak{s}_f''(u))^2 \,\mathrm{d}u \,\mathrm{d}t \,\mathrm{d}x \\
&\leq \frac{h^4}{128} \int_0^{\frac{h}{2}} (f''(u) - \mathfrak{s}_f''(u))^2 \,\mathrm{d}u \\
&\leq \frac{h^4}{128} \int_0^{\frac{h}{2}} (f''(u))^2 \,\mathrm{d}u,
\end{aligned}$$

and once again by symmetry this extends to $[0, h]$ and then to $[0, 1]$, yielding the desired result. Assume now that $f \in H^4(0, 1)$. These proofs are essentially the same, without needing the orthogonality property because $\mathfrak{s}_f^{(4)} = 0$. In the following proof, we use the fact that there exist $\theta_2, \theta_3 \in [0, h]$ such that $f^{(2)}(\theta_2) - \mathfrak{s}_f^{(2)}(\theta_2) = f^{(3)}(\theta_3) - \mathfrak{s}_f^{(3)}(\theta_3) = 0$, which can be proved using Rolle's theorem four times on $f - \mathfrak{s}_f$ (1 time), $f' - \mathfrak{s}_f'$ (2 times) and $f^{(2)} - \mathfrak{s}_f^{(2)}$ (1 time), respectively.

$$\begin{aligned}
\int_0^{\frac{h}{2}} \left((f - \mathfrak{s}_f)^{(3)}(x) \right)^2 \mathrm{d}x &= \int_0^{\frac{h}{2}} \left(\int_{\theta_3}^x (f - \mathfrak{s}_f)^{(4)}(t) \,\mathrm{d}t \right)^2 \mathrm{d}x \\
&\leq \int_0^{\frac{h}{2}} |x - \theta_3| \int_{\theta_3}^x (f^{(4)}(t))^2 \,\mathrm{d}t \,\mathrm{d}x \\
&\leq \frac{3h^2}{8} \int_0^h (f^{(4)}(t))^2 \,\mathrm{d}t,
\end{aligned}$$

since the worst case is $\theta_3 = h$, which yields $C_3 = \frac{\sqrt{3}}{2}$. Then,

$$\begin{aligned}
\int_0^{\frac{h}{2}} \left((f - \mathfrak{s}_f)^{(2)}(x) \right)^2 \mathrm{d}x &= \int_0^{\frac{h}{2}} \left(\int_{\theta_2}^x \int_{\theta_3}^t (f - \mathfrak{s}_f)^{(4)}(u) \,\mathrm{d}u \,\mathrm{d}t \right)^2 \mathrm{d}x \\
&\leq \int_0^{\frac{h}{2}} |x - \theta_2| \int_{\theta_2}^x |t - \theta_3| \int_{\theta_3}^t (f^{(4)}(u))^2 \,\mathrm{d}u \,\mathrm{d}t \,\mathrm{d}x \\
&\leq \int_0^{\frac{h}{2}} |x - \theta_2| \,\mathrm{d}x \int_0^h |t - \theta_3| \,\mathrm{d}t \int_0^h (f^{(4)}(u))^2 \,\mathrm{d}u \\
&\leq \frac{3h^4}{16} \int_0^h (f^{(4)}(u))^2 \,\mathrm{d}u,
\end{aligned}$$

which yields $C_2 = \frac{\sqrt{201}}{24}$. Then,

$$\begin{aligned}\int_0^{\frac{h}{2}} \left((f-\mathfrak{s}_f)^{(1)}(x)\right)^2 \mathrm{d}x &= \int_0^{\frac{h}{2}} \left(\int_0^x \int_{\theta_2}^t \int_{\theta_3}^u (f-\mathfrak{s}_f)^{(4)}(v)\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t\right)^2 \mathrm{d}x \\ &\leq \int_0^{\frac{h}{2}} x \int_0^x |t-\theta_2| \int_{\theta_2}^t |u-\theta_3| \int_{\theta_3}^u (f^{(4)}(v))^2\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}x \\ &\leq \int_0^{\frac{h}{2}} x(h^2-x^2)\,\mathrm{d}x \int_0^h |u-\theta_3|\,\mathrm{d}u \int_0^h (f^{(4)}(v))^2\,\mathrm{d}v \\ &\leq \frac{7h^6}{256}\int_0^h (f^{(4)}(v))^2\,\mathrm{d}v,\end{aligned}$$

which yields $C_1 = \frac{\sqrt{7}}{8\sqrt{2}}$. And finally,

$$\begin{aligned}\int_0^{\frac{h}{2}} ((f-\mathfrak{s}_f)(x))^2\,\mathrm{d}x &= \int_0^{\frac{h}{2}} \left(\int_0^x \int_0^t \int_{\theta_2}^u \int_{\theta_3}^w (f-\mathfrak{s}_f)^{(4)}(w)\,\mathrm{d}w\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t\right)^2 \mathrm{d}x \\ &\leq \int_0^{\frac{h}{2}} x \int_0^x t \int_0^t |u-\theta_2| \int_{\theta_2}^u |v-\theta_3| \int_{\theta_3}^v (f^{(4)}(w))^2\,\mathrm{d}w\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}x \\ &\leq \int_0^{\frac{h}{2}} x \int_0^x t \int_0^t |u-\theta_2| \int_{\theta_2}^u |v-\theta_3|\,\mathrm{d}v\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}x \int_0^h (f^{(4)}(w))^2\,\mathrm{d}w \\ &\leq \int_0^{\frac{h}{2}} x \int_0^x t \int_0^t (h-u)\,\mathrm{d}u\,\mathrm{d}t\,\mathrm{d}x \int_0^h |v-\theta_3|\,\mathrm{d}v \int_0^h (f^{(4)}(w))^2\,\mathrm{d}w \\ &\leq \frac{9h^8}{10240}\int_0^h (f^{(4)}(v))^2\,\mathrm{d}v,\end{aligned}$$

which yields $C_0 = \frac{3\sqrt{5}}{160}$. □

As we have mentionned before, those estimates are useful only if cubic splines are used to approximate more regular functions. This is the prurpose of the next section, which introduces a space of regular heptic splines.

I.3.2.d Heptic splines

In this section we introducte heptic splines – polynomials of degree 7 by part. These can be used as an approximation method, or more simply as an interpolation method. We shall only present here a basis of the heptic splines that are in $H^4(0,1)$. As for cubic splines, we shall define the space of heptic splines as

$$\$_7^h = \left\{\mathfrak{s} : [0,1] \to \mathbb{R}, \quad \forall\, i \in \{0,\dots,N-1\}, \mathfrak{s}_{|[x_i,x_{i+1}]} \in \mathbb{R}_7([x_i,x_{i+1}])\right\}, \tag{I.3.68}$$

as well as $\mathcal{S}_7^h$ the space of three times differentiable heptic splines. We introduce the following basis of $\mathcal{S}_7^h$: for $i \in \{0,\dots,N\}$,

$$\xi_i^0 \,:\, x \mapsto \begin{cases} \frac{1}{h^7}(x-x_{i-1})^4\left(-20(x-x_i)^3+10h(x-x_i)^2-4h^2(x-x_i)+h^3\right) & \text{if } x_{i-1} \leq x \leq x_i \\ \frac{1}{h^7}(x-x_{i+1})^4\left(20(x-x_i)^3+10h(x-x_i)^2+4h^2(x-x_i)+h^3\right) & \text{if } x_i \leq x \leq x_{i+1} \end{cases}$$

$$\xi_i^1 \,:\, x \mapsto \begin{cases} \frac{1}{h^6}(x-x_{i-1})^4(x-x_i)\left(10(x-x_i)^2-4h(x-x_i)+h^2\right) & \text{if } x_{i-1} \leq x \leq x_i \\ \frac{1}{h^6}(x-x_{i+1})^4(x-x_i)\left(10(x-x_i)^2+4h(x-x_i)+h^2\right) & \text{if } x_i \leq x \leq x_{i+1} \end{cases}$$

$$\xi_i^3 \,:\, x \mapsto \begin{cases} \frac{1}{2h^5}(x-x_{i-1})^4(x-x_i)^2\left(-4(x-x_i)+h\right) & \text{if } x_{i-1} \leq x \leq x_i \\ \frac{1}{2h^5}(x-x_{i+1})^4(x-x_i)^2\left(4(x-x_i)+h\right) & \text{if } x_i \leq x \leq x_{i+1} \end{cases}$$

$$\xi_i^3 \,:\, x \mapsto \begin{cases} \frac{1}{6h^4}(x-x_{i-1})^4(x-x_i)^3 & \text{if } x_{i-1} \leq x \leq x_i \\ \frac{1}{6h^4}(x-x_{i+1})^4(x-x_i)^3 & \text{if } x_i \leq x \leq x_{i+1} \end{cases}$$

and furthermore every basis function ξ_i^j is equal to zero outside of $[x_{i-1}, x_{i+1}]$ – see Figure I.5 for an illustration of some elements of the basis. This basis has been designed so that ξ_i^k defines the value of the kth derivative at point x_i:

$$\forall\,(i,j) \in \{0,\dots,N\}^2,\ \forall\,(k,l) \in \{0,1,2,3\}^2, \qquad (\xi_i^k)^{(l)}(x_j) = \begin{cases} 1 \text{ if } (i,k) = (j,l) \\ 0 \text{ otherwise.} \end{cases} \tag{I.3.69}$$

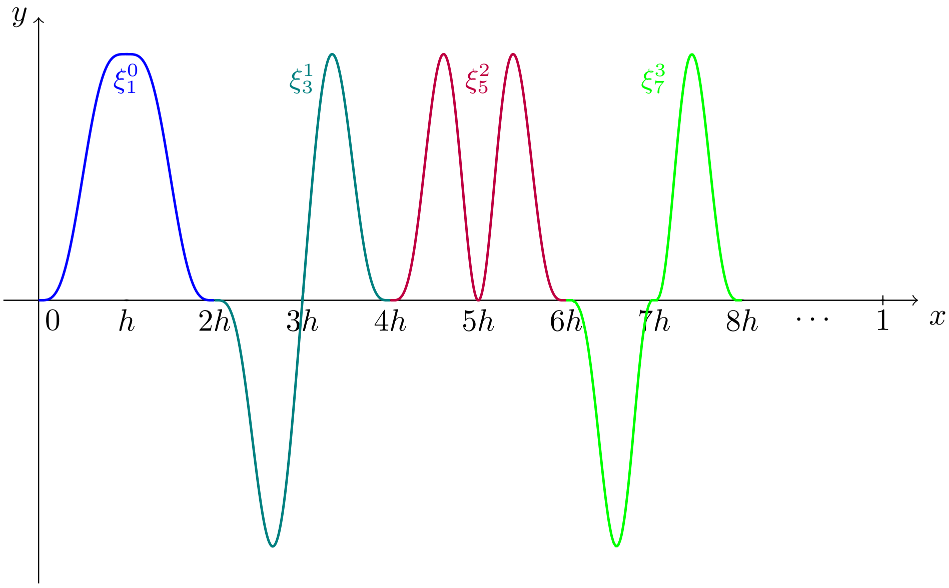


Figure I.5: Heptic spline basis.

Using this basis, one could interpolate cubic splines in an optimal manner (recall Theorem I.53). However, this is rather tedious (one could even say horridly so) and would only be worth it if one has already worked up the necessary discretisation error bounds to use them – for more details, we refer the reader to Remark III.14.

I.4 Rigorous numerics

In this section, we shall introduce the field of rigorous numerics, its goals and some of its techniques, then will elaborate on how we use it in this thesis, giving details about the core difficulties induced and how we tackle them.

I.4.1 Introduction to rigorous numerics

The objective of this section is to introduce the field of rigorous numerics. Rigorous numerics aim at controlling the result of numerical computations, by guaranteeing that the result the user wished to compute is included in a computed set. In particular, it is designed to precisely bound round-off errors.

Recall that when numerically modelling a phenomena, different types of errors can arise:

- firstly, the modelling error, which is the difference between reality and the model. It is the most difficult to characterise and even more to estimate.
- secondly, the discretisation error: when modelling a phenomenon depending on space or time, discretisation is very often needed to numerically solve the model. When modelling using differential equations, the errors are estimated and bounded using tools from the field of *Numerical Analysis*. See Section I.3 for some details about how we use it in this thesis.

- last but not least, when computing discretisation schemes on standard digital computers, approximations are made at each computation because of the finite-byte representation of numbers. This is the primary goal of *Rigorous numerics*, which we shall present in this section.

Rigorous numerics, also known as validated numerics, verified computation or any combination of these words, is using set-based arithmetic to bound and guarantee that the desired result is included in the returned set. Although it has similar goals to the fields of *symbolic computation* or *proof assistants*, its methods are fundamentally different:

- rigorous numerics aims at providing error bounds for computations that would usually be done with floating-point arithmetic, prone to round-off errors
- symbolic computation deals with mathematical expressions, storing variables not given any value, and is primarily meant to automatise complex routines to obtain closed formulae
- proof assistants are tools designed to assist the verification of formal proofs, checking that every hypotheses has been formulated or verified, and automatically providing proofs of simple statements.

Rigorous numerics have been developped since the 80s, and now several toolboxes are available for everyone to use: in MATLAB/Octave (`INTLAB`), in C++ (`kv`, `CAPD` and `Boost Safe Numerics`), in C (`Arb`) or Julia (`JuliaIntervals`). In this thesis, the rigorous numerics have been implemented using `INTLAB` [Rum99], developped and maintained by Prof. Dr. Siegfried M. Rump. It includes different methods that have been developped over the years to provide tight estimates of various verified computations, including *interval arithmetic* and *affine arithmetic*.

Interval arithmetic is the simplest: each variable is represented by an interval or a cartesian product of intervals. It has been extensively studied and thouroughly encoded, and is the most used method because of its simple representation, which allows for easier bounding processes, and its relatively fast computations. For a well-detailed introduction to interval arithmetic, see [Tuc11].

On the other hand, *affine arithmetic* represents data as zonotopes – centrally symmetric polytopes – stored as a centre position plus error directions and distances. This much more general class of convex sets, and thus allows for tighter error bounds. The price to pay lies firstly in the theoretical complexity of computations of certified simple operations on the sets and secondly in the numerical computation time: for example, compared to floating-point arithmetic, a simple 100×100 matrix-matrix product can take twice as long with interval arithmetic, and a thousand times longer on affine arithmetic.

Apart from the computation time, the main hindrance for rigorous numerics is the so-called *wrapping effect*: if not treated carefully, the radii of the considered sets can increase exponentially with each computation. This is especially true for interval arithmetic, due to the rigidity of its representation of errors. Consider for example a 2D rotation matrix A_θ of angle $\theta = 1$. The computation of $\cos(\theta)$ and $\sin(\theta)$ is done up to double precision, so with interval arithmetic their corresponding intervals have radii of approximately $2.23 \cdot 10^{-16}$. Consider the three following algorithms to compute A_θ^{1024}:

1. Store $B = \mathrm{Id}_2$. Then do 1024 times $B = A_\theta B$ and return B
2. Store $B = A_\theta$. Then do 10 times $B = B^2$ and return B
3. Compute $\theta' = 1024\,\theta$, and return $A_{\theta'}$.

The result of those three algorithms in interval arithmetic have respectively radii of $2.9 \cdot 10^{125}$, $3.4 \cdot 10^{-12}$ and $2.3 \cdot 10^{-16}$. The main reason for this staggering error lies in the storing of each iterate as a box centred in 0: at each iteration, the 4D box that represents the matrix is rotated by θ, and then upper-bounded by another non-rotated box, with greater radii. Therefore, the total error increases exponentially with the number of iterations.

The wrapping effect is much less pronounced in affine arithmetic: the result of the three algorithms in affine arithmetic returns raddi of $5.2 \cdot 10^{-13}$, $5.0 \cdot 10^{-13}$ and $2.3 \cdot 10^{-16}$. The two first algorithms yield very similar thanks to the ability of affine arithmetic to accurately capture rotations. The result of the third algorithm is identical whether computed with interval or affine arithmetic because all computations are done in $\mathbb{R}$, where they are equivalent.

Despite this obvious advantage of affine arithmetic, its computation time often proves prohibitive. In this thesis, we shall hence use interval arithmetic. In the next section, we present the carefully tuned algorithms used to encode the functionals needed in this thesis, aiming to minimise both wrapping effect and computation time. Everything is written to match the notations and context of Chapter III, since it is the most computationnally heavy, but every method should be applied in all settings.

I.4.2 Computation of the main functionals

Recall that most of the results of this thesis rely on the minimisation, computation and verification of the functional

$$J : p_f \mapsto \int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f)\,\mathrm{d}t - \langle y_f - S_T y_0, p_f \rangle, \tag{I.4.1}$$

which is discretised in time and space by

$$J_{\Delta t,h} : p_{fh} \mapsto \Delta t \sum_{k=0}^{N_0-1} \sigma_{B\mathcal{U}}((\mathrm{Id} + \Delta t A_h)^{-k} p_{fh}) - \langle y_f, p_{fh} \rangle + \langle y_0, (\mathrm{Id} + \Delta t A_h)^{-N_0} p_{fh} \rangle. \tag{I.4.2}$$

The very first to note is that minimising this functional does not require the use of rigorous numerics. The sole purpose of minimisation is to provide a minimiser p_{fh}, that will be an appoximation of the true minimiser (if it exists), but ultimately the functional $J_{\Delta t,h}$ will be computed at point p_{fh} and the interval containing $J(p_f)$ will be computed the same whether the minimisation was done with or without numerical verification. Therefore, we shall only focus on verifying the computation of $J_{\Delta t,h}$ and its main components: an exponential of matrices, discretisation of a time integral, a support function and inner products.

Here, recall that we consider a finite-dimensional subspace V_h of X, V_h being generated by a given basis $(\varphi_i)_i$. Therefore in this paragraph we consider that every variable $p \in V_h$ is identified with the vector of $\mathbb{R}^{\dim V_h}$ corresponding to its representation in the $(\varphi_i)_i$ basis. An interval $\mathcal{Y}$ of V_h is hence a closed (potentially unbounded) hypercube of $\mathbb{R}^{\dim V_h}$ defined with its two opposite vertices $\underline{y}, \overline{y} \in \mathcal{Y}$ such that for all $y \in \mathcal{Y}$, for all $i \in \{1, \dots, \dim V_h\}$, $\underline{y}_i \leq y_i \leq \overline{y}_i$.

Exponential of matrices. Along this thesis, all exponential of matrices have been computed using Euler schemes, whether explicit II.3.2 or implicit II.3.3, III.2.2. We shall focus on the implicit Euler scheme, since numerically it has one more component: inversion of a matrix. Consider for this paragraph a matrix C, and assume you wish to compute $C^{-k} p_{fh}$ for $k \in \{0, \dots, N_0\}$. For now, let us fix k.

Three algorithms might come to mind when computing $p_k = C^{-k} p_{fh}$:

1. iterately solve $C p_{i+1} = p_i$, where $p_0 = p_{fh}$
2. compute C^k and then solve $C^k p_k = p_{fh}$
3. compute C^{-1}, then compute $(C^{-1})^k$ and finally $C^{-k} p_{fh}$

The first two algorithms will return high round-off errors: the first one because of the wrapping effect illustrated earlier – a good rule of thumb with interval arithmetic is to avoid loops or repeated operations on a single variable. The second algorithm will fail for a different reason: even if computed smartly, C^k will have bigger errors than C. This is due to the nature of the problems we study: essentially C has non-negative eigenvalues, and thus C^k and its round-off errors will get "bigger" with k. Ultimately, the risk is that C^k will get "closer to 0", which will make the errors explode when inverting C^k.

The third algorithm, however, will take advantage that C^{-1} will have non-positive eigenvalues, so that C^{-k} and its round-off errors will decrease with k. As for the computation of C^{-k} from C^{-1}, one should consider a *fastexp* technique, that is, computing the C^{-2^i} with $i \in \{1, \dots, \log_2(k)\}$ and deducing C^{-k} from the decomposition in base 2 of k. This furthermore allows easy vectorisation to compute all $\{C^{-k}p_{fh},\, k \in \{0, \dots, N_0\}\}$.

Time integral. Assuming readily computable support function and exponentials, the computation of an approximation of the time-integral seems easy, especially since it is simply approximated using the rectangle rule. However, when dealing with very precise discretisation parameters (sometimes $1,000,000$ time-meshes and $2,000$ space-meshes), a memory-efficient algorithm becomes essential to drastically reduce computation time. In order to do that, one should split the computation into manageable batches. Instead of storing all intermediate results, process the time steps in blocks that fit into the available memory cache. For each batch, compute the required exponentials and support functions, accumulate the partial sums and its round-off errors, and then proceed to the next batch. This approach minimizes memory usage and leverages cache locality, significantly speeding up the computation.

For example, in our implementation, processing the entire space-time grid at once would require an unmanageable amount of memory and could take years to complete. By dividing the computation into batches (e.g., 10^4 time steps per batch), we once reduced the total computation time from an estimated 19 years to about 1 hour on a standard workstation. This batching strategy is crucial for handling large-scale problems efficiently with rigorous numerics.

Support function. Another thing to consider when dealing with interval arithmetic is that although most common routines have been efficiently encoded, some specific functions have to be carefully thought about. This is the case of support functions: here we are concerned with the support function of $B\mathcal{U}$. We thus have

$$\sigma_{B\mathcal{U}}(y) = \sup_{u\in\mathcal{U}} \langle Bu, y\rangle. \tag{I.4.3}$$

When considering $\mathcal{Y} \subset X$ an interval centred on $y^0 \in X$, we have

$$\sigma_{B\mathcal{U}}(y^0) \in \sigma_{B\mathcal{U}}(\mathcal{Y}) := \left[\inf_{y\in\mathcal{Y}} \sigma_{B\mathcal{U}}(y),\, \sup_{y\in\mathcal{Y}} \sigma_{B\mathcal{U}}(y)\right], \tag{I.4.4}$$

and both those bounds have to be computed, or at least lower- and upper-bounded. In particular, the supremum is reached on an extreme points of $\mathcal{Y}$ – which are known, since $\mathcal{Y}$ is an interval – and the infimum can be tackled with standard optimisation techniques. Of course, this is highly dependent on the setting, so let us give an example in the case of internal control of the heat equation considered in III.3.3.a. In this context $\mathcal{Y} \subset V_h$, where V_h is the set of finite elements on a regular mesh of $[0, 1]$ and satisfying zero boundary conditions, so

$$\mathcal{Y} = \{y \in V_h, \forall\, x_i = ih, y(x_i) \in [\underline{y_i}, \overline{y_i}]\}. \tag{I.4.5}$$

Set $B = \mathrm{Id}$ and $\mathcal{U} = \{x \in L^2(0, 1), \|x\|_2 \leq 1\}$. In this simple case, we have that for all $y \in L^2(0, 1)$, $\sigma_{B\mathcal{U}}(y) = \|y\|_2$, and thus

$$\sigma_{B\mathcal{U}}(y^0) \in \sigma_{B\mathcal{U}}(\mathcal{Y}) = \left[\inf_{y\in\mathcal{Y}} \|y\|_2,\, \sup_{y\in\mathcal{Y}} \|y\|_2\right]. \tag{I.4.6}$$

For $\mathcal{Y} \subset V_h$, it is easily computed that

$$\inf_{y\in\mathcal{Y}} \|y\|_2 = \|y^\star\|, \quad \text{where} \quad y^\star \in \mathcal{Y} \text{ and on the mesh } y^\star(x_i) = \inf_{y\in[\underline{y_i},\overline{y_i}]} |y|. \tag{I.4.7}$$

However, the supremum has no closed form expression, aside from the cumbersome

$$\sup_{y\in\mathcal{Y}} \|y\|_2 = \max\left\{\|y\|_2,\ y\in V_h,\ \forall\, i,\ y(x_i)\in\{\underline{y_i},\overline{y_i}\}\right\}, \tag{I.4.8}$$

the right-hand set being of cardinal $2^{\dim V_h}$. One way to efficiently and closely upper-bound is to separate on each interval of the mesh:

$$\sup_{y\in\mathcal{Y}} \|y\|_2^2 \leq \sum_{i=0}^{\dim V_h - 1} \max\left\{\int_{x_i}^{x_{i+1}} y^2(x)\,\mathrm{d}x,\ y\in V_h,\ y(x_i)\in\{\underline{y_i},\overline{y_i}\},\ y(x_{i+1})\in\{\underline{y_{i+1}},\overline{y_{i+1}}\}\right\}. \tag{I.4.9}$$

In the various examples in Subsection III.3.3.a, these are the upper and lower-bounds we use to certify the support function of the reachable set. This example, although extremely simple at first glance, showcases the difficulties inherent to any encoding using rigorous numerics. When considering more elaborate examples, one also has to consider the encoding of B: for example, if $B = \chi_\omega$, precautions have to be taken if the boundaries of ω do not coincide with the mesh associated to V_h.

Scalar products and norm of A^*p_f. One final point to mention is the computation of the inner products : $\langle y_f, p_f\rangle$ and $\langle y_0, S_T^* p_f\rangle$, as well as the norm $\|A^*p_f\|$. In order to compute and certify those elements, it is assumed in Subsection III.2.4 to have closed formulae of $\langle y_f, p_{fh}\rangle$ and $\langle y_0, p_{fh}\rangle$ for every $p_{fh}\in V_h$. If however one has access to closed formulae of $\langle y_f, p_f\rangle$ directly, given $p_f\in\mathcal{D}(A^*)$ an interpolation of p_{fh}, then of course one should compute this instead, and forego the discretisation error associated to this inner product. However, as has been noted throughout this section, one has to be careful that the round-off errors generated by the computation of $\langle y_f, p_f\rangle$: should they exceed the discretisation error associated to the computation of $\langle y_f, p_{fh}\rangle$, the first method should be favoured – this can in particular happen for a very fine discretisation: the discretisation error decreases with the discretisation parameters, whereas the round-off error increases with the dimension of the considered vectors, and is all the more present when computing the complicated $\langle y_f, p_f\rangle$ than the simple $\langle y_f, p_{fh}\rangle$.

On another note, the computation of $\|A^*p_f\|$, which is essential to bound the discretisation errors of the whole functional, needs to be certified. Although it might be possible to compute $\|A_h^*p_{fh}\|$ and bound its discretisation and round-off errors, in the settings addressed in this thesis – where interpolation is performed using cubic splines – a closed formula is preferable. Here as well, remark that increasing the dimension of V_h, and thus complicating the formula of the interpolation p_f will increase the round-off errors on $\|A^*p_f\|$. This forms a seemingly unavoidable barrier to computer-assisted proofs involving discretisation of partial differential equations: no matter how well you discretise or how finely you tune your rigorous numerics, there may be some true yet unprovable results.

II

Certified non-reachability for finite-dimensional control problems

Contents

Abstract

It is customary to design a control system in such a way that, whatever the chosen control satisfying the constraints, the system does not enter so-called unsafe regions. This work introduces a general computer-assisted methodology to prove that a given linear control system with compact constraints avoids a chosen unsafe set at a chosen final time T. Relying on support hyperplanes, we devise a functional such that the property of interest is equivalent to finding a point at which the functional is negative. Actually evaluating the functional first requires time-discretisation. We thus provide explicit, fine discretisation estimates for various types of matrices underlying the control problem. Second, computations lead to roundoff errors, which are dealt with by means of interval arithmetic. The control of both error types then leads to rigorous, computer-assisted proofs of non-reachability of the unsafe set. We illustrate

the applicability and flexibility of our method in different contexts featuring various control constraints, unsafe sets, types of matrices and problem dimensions.

The following article is a nearly identical transcript of the paper *Computer-assisted proofs of non-reachability for finite-dimensional linear control systems* co-written with Camille Pouchol, Yannick Privat and Christophe Zhang [Has+24], accepted in *SIAM Journal of Control and Optimisation* in May 2025.

II.1 Introduction

This article is dedicated to the rigorous study of non-reachable states of a constrained controlled linear system. More precisely, we are interested in guaranteeing that, at a given time $T > 0$, the control system cannot enter a prescribed *unsafe region*, whatever the choice of control satisfying the given constraints.

We consider the linear autonomous (time invariant) control system

$$\begin{cases} y'(t) = Ay(t) + Bu(t), \\ y(0) = y_0, \end{cases} \tag{$\mathcal{S}$}$$

where $y_0 \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$.

Given $y_0 \in \mathbb{R}^n$, a closed convex set $\mathcal{Y}_f \subset \mathbb{R}^n$, a time horizon $T > 0$ and a compact set $\mathcal{U} \subset \mathbb{R}^m$, we investigate the ($\mathcal{U}$-)constrained reachability problem, i.e., the problem of determining if there exists a control u such that the solution to ($\mathcal{S}$) with control u satisfies $y(T) \in \mathcal{Y}_f$, under the additional constraint that $u(t) \in \mathcal{U}$ for a.e. $t \in (0, T)$. If such a control exists, we shall say that *$\mathcal{Y}_f$ is $\mathcal{U}$-reachable from y_0 in time T.*

Our aim is to develop a general, flexible and certifiable methodology, resting on numerical computations, to show that $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time T. Ultimately, the interested user should be able to provide all parameters A, B, T, y_0, $\mathcal{U}$ and $\mathcal{Y}_f$ and, whenever it is the case, be returned the mathematically certified assertion that $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time T.

II.1.1 Methodology: non-reachability criterion and certification issues

Support functions. Throughout, finite-dimensional spaces $\mathbb{R}^n$, $\mathbb{R}^m$ will be endowed with the standard Euclidean inner products. If we have $C \subset H$ with H a Hilbert space, σ_C will denote the *support function* of C defined by

$$\forall\, x \in C, \quad \sigma_C(x) := \sup_{y\in C} \langle x, y\rangle.$$

Non-reachability by separation. By means of separating hyperplanes, we will establish a necessary and sufficient criterion for non-reachability, involving a suitably defined function $J : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, in the following form:

$$(\exists\, p_f \in \mathbb{R}^n, \quad J(p_f) < 0) \qquad \Longleftrightarrow \qquad \mathcal{Y}_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T. \tag{II.1.1}$$

The precise definition of J (together with Figure II.1 to convey the corresponding intuition) will be given in Section II.2.1, and involves the support functions $\sigma_{\mathcal{U}}$ and $\sigma_{\mathcal{Y}_f}$, which we assume to be known explicitly.

The proof of (II.1.1) is the object of Proposition II.2. In the case where $p_f \in \mathbb{R}^n$ such that $J(p_f) < 0$ is found, we will say that p_f is a *dual certificate* (that $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time T). One should note that such a dual certificate only proves the non-reachability at time T, and not for all $t \in [0, T]$. Sufficient conditions for the non-reachability at all times $t \in [0, T]$ are proposed in Proposition II.7 and the following remark.

Computer-assisted proof of non-reachability. In what follows, we will exploit this criterion by producing vectors that satisfy it numerically. This raises questions pertaining to the error propagation inherent to every numerical method.

> *Certified approach for non-reachability.*
>
> - How can one evaluate the functional J, in order to exhibit an element $p_f \in \mathbb{R}^n$ satisfying property (II.1.1) numerically?
> - How can one then **certify** the numerical result, which implies non-reachability? That is, guarantee that it is not flawed by various numerical approximations?

In order to carry out these two steps, there will in turn be two main difficulties.

(i) We will not have access to J but only to proxies obtained by discretisation, which we generically denote J_d. Indeed, the definition of J involves a time-integral, as well as the solution to a linear ODE involving A^* (which amounts to computing the matrix exponentials $t \mapsto e^{tA^*}$). When these are not known explicitly, we will resort to simple time discretisation schemes (implicit Euler, etc) and provide a **bound on the error in terms of discretisation parameters**. One key aspect of our approach is that these bounds must be derived with explicit constants.

(ii) All computations will lead to **round-off errors**, which must be accounted for. To that end, we will use a MATLAB/Octave toolbox called INTLAB (INTerval LABoratory) [Rum99]. This code is an interval arithmetic library, entirely written in MATLAB. It provides tools for performing numerical computations with arbitrary-precision arithmetic.

All in all, if for a given $p_f \in \mathbb{R}^n$ one lets $E_\mathrm{d}(p_f)$ (for the discretisation errors) and $E_\mathrm{r}(p_f)$ (for the round-off errors), we will have

$$J(p_f) \in [J_\mathrm{d}(p_f) - E_\mathrm{d}(p_f) - E_\mathrm{r}(p_f), J_\mathrm{d}(p_f) + E_\mathrm{d}(p_f) + E_\mathrm{r}(p_f)]. \tag{II.1.2}$$

Hence, we will take advantage of the fact that if

$$J_\mathrm{d}(p_f) + E_\mathrm{d}(p_f) + E_\mathrm{r}(p_f) < 0,$$

then y_f is not $\mathcal{U}$-reachable from y_0 in time T.

Here we stress that the notion of **certification** we are concerned with has to do with the numerical part of our work. The starting point of this work is a theoretical necessary and sufficient condition for non-reachability. For a given system, we can determine whether it is satisfied numerically. Certifying this part then makes this numerical result theoretically sound, thus producing a **computer-assisted proof**. Another key aspect of our methodology is to return a dual certificate p_f that *certifies* the corresponding mathematical statement: consequently, any user with access to their own discretised version of the functional J with corresponding error estimates, can verify the result upon using interval arithmetic.

II.1.2 State of the art & connections to existing results

The notion of constraint-free controllability of autonomous linear systems dates back to Kalman's seminal works. Its generalisation to infinite-dimensional systems is more recent. For further details on these concepts, we refer the reader to the review books [Lio92; J-M07]. Since the 70's, but more specifically in recent years, several works have investigated the addition of further constraints, satisfied whether by the control itself, or by the controlled trajectory.

Some of these works are theoretical in nature, with a focus on unbounded constraints. Particular interest has been given to the problems of exact controllability by positive controls due to their physical relevance [Bra72; Res05; FHL92; Kla96; PZ18; PZ19; LM21]. Attention was also paid to adding constraints on the controlled trajectory [LTZ18; LTZ21; Erv20]. Unbounded (sparsity) constraints have also been considered [Zua10; PTZ24].

In this article, we focus on the implementation of a method to numerically certify that a set of unsafe states is unreachable at a given time $T > 0$, for compact constraint sets on the control. Our approach is specific to autonomous (time invariant) linear systems. Regarding more general dynamical systems, closely related questions have been addressed in the past: for instance, how to numerically approximate the reachable set at time T, or guarantee that computed trajectories will not meet the given unsafe set at any time $t > 0$.

In finite dimension, several methods have been elaborated to provide approximations of the reachable set (for example, see the recent survey [AFG21]): among others, let us mention the use of Hamilton-Jacobi type equations [MBT05; CT18], the design of barrier functions for trajectories to avoid unsafe regions [PJ04; KBH18], and set propagation [AFG21]. Let us roughly describe each of these approaches.

In [MBT05; CT18], a backwards reachable set is characterised as the zero sublevel set of the viscosity solution to a Hamilton-Jacobi type partial differential equation, with important applications to the safety of automated systems. This is formally related to our approach, as we also characterise non-reachability by the existence of negative values for a certain numerical criterion. As we will see, in this paper the convexity of the reachable set and the linearity of the system allow us to exploit this characterisation to produce numerical certificates of non-reachability.

In [PJ04; KBH18], the authors introduce the notion of barrier functions, appropriately defined from the system dynamics to ensure that trajectories do not enter an unsafe zone. An important element of these methods is that these certificates are valid for all positive times $t > 0$, a very strong property which is not required in other methods, and in particular in this article. Moreover, the computation of barrier certificates for a given system remains a challenging problem, both theoretically and numerically.

Set propagation is a class of methods for computing a guaranteed over-approximation or under-approximation of the reachable set of continuous systems. Starting from the set of initial states, the idea is to iteratively and adequately propagate a sequence of sets according to the system dynamics [GLM06], which are guaranteed to contain, or be contained in, the reachable set. Such an algorithm has been developed in [LG10; LG09] for finite-dimensional compact convex constraints. An important hurdle is then the so-called *wrapping effect*, which is the accumulation of computational errors. The crux of set propagation techniques is to circumvent this difficulty by using appropriate propagation formulae. In this article, the wrapping effect is avoided using duality and considering the solution to a single backward equation.

Separation arguments, as used in this article, already appear in reachability analysis [KV02a; KV02b; LG10; LG09; Bai+07]. However, an important contribution we make is recasting it in terms of the sign of the function J, in such a way that interval arithmetic can be applied to certify the end result – a feature which seldom appears in the literature.

Using our approach, one can prove that the reachable set is contained in a half-space. Computing several such half-spaces allows for the creation of a bounded convex polytope guaranteed to contain the reachable set, but this can quickly become computationally expensive, especially as the dimension of the problem increases. There exist other ways to over- or under-approximate reachable sets, which rely on geometric properties. In the special case of ellipsoidal constraint sets, we refer the reader to [KV02a; KV02b]. More generally, for compact convex constraints, the reachable set can be approximated from the outside using support functions [Bai+07].

While the above-mentioned works provide theoretical criteria for finite dimensions, the case of infinite dimensions remains largely open.

For a more comprehensive review of the literature, we refer the reader to Section 0.3.

Extensions and perspectives. We make the assumption that the support functions $\sigma_{\mathcal{U}}$ and $\sigma_{\mathcal{Y}_f}$ are known exactly. If it were not the case, our approach could be extended, provided that one has a procedure to numerically evaluate them, together with a way to control the corresponding error.

The approach we have developed can be adapted *mutatis mutandis* to non-homogeneous non-autonomous linear systems of the form $x'(t) = A(t)x(t) + B(t)u(t)+v(t)$, for some $v \in L^2(0, T; \mathbb{R}^n)$. The price to pay lies in the error formulae, where the exponential matrix e^{tA} is replaced by the resolvent associated with the function $A(\cdot)$. The resulting formulae would then be slightly less accurate than those we obtained.

Another relevant issue would concern non-reachability in fixed time for non-linear control systems under control sampling; see e.g. [BT21]. This amounts to imposing specific constraints on the control, assuming it to be piecewise constant with values in a given prescribed set. In the case of a linear system, our approach would apply provided that one provides efficient ways to compute or approximate the support function of these particular types of constraint sets.

In the same vein, the reachability criterion can be extended without effort to Hilbert spaces (see for instance [LY12] for infinite dimensional time optimal control problems). This is why we expect our method to accommodate **infinite-dimensional linear control systems**, provided that the space discretisation errors are also estimated. This will be the subject of further work, focusing in particular on the heat equation.

Our work adresses non-reachability. The natural complementary question is that of reachability: can one provide certified methods to show that a target y_f (or more generally, a set $\mathcal{Y}_f$) is reachable? We intend to tackle this problem as well, using similar geometric ideas.

Finally, a more prospective research direction is to investigate generalisations to non-linear control systems. It is obvious that the methodology will have to be thoroughly modified, since our approach fundamentally rests on the linearity of L_T.

Outline of the article. In Section II.2, we introduce the criterion J and specify the separation argument, which allows us to recast the non-reachability property. Section II.3 focuses on numerical methods for calculating J, using several possible discrete versions. Their relevance is discussed based on the available information about A and its matrix exponential, and in each case, we provide fully explicit error bounds. Finally, the Section II.4 is entirely devoted to numerical experiments. After specifying the methodology leading to computer-assisted proofs of non-reachability, we apply it to three examples, with variable dimensions and constraint sets. We present concrete statements, each rigorously proven using our computer-assisted methodology.

II.2 Non-reachability by separation

II.2.1 Main result

Consider the linear autonomous control system

$$\begin{cases} y'(t) = Ay(t) + Bu(t), & t \in [0,T], \\ y(0) = y_0, \end{cases} \tag{$\mathcal{S}$}$$

where $y_0 \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, with $u \in E := L^2(0,T;\mathbb{R}^m)$. Recall that we make the following assumption regarding the constraint set $\mathcal{U}$ and the unsafe set $\mathcal{Y}_f$:

$$\mathcal{U} \text{ is compact,} \qquad \mathcal{Y}_f \text{ is closed and convex.} \tag{H}$$

The solution to ($\mathcal{S}$) at the final time T is characterised by Duhamel's formula and is written as:

$$y(T) = e^{TA}y_0 + L_T u, \qquad \text{where} \qquad L_T u := \int_0^T e^{(T-t)A}Bu(t)\,dt.$$

Letting $L(H_1, H_2)$ denote the set of linear continuous operators between two Hilbert spaces H_1 and H_2, it is standard that L_T defines an operator in $L(E,\mathbb{R}^n)$. Its adjoint, $L_T^* \in L(\mathbb{R}^n, E)$, is defined for $p_f \in \mathbb{R}^n$ by $L_T^* p_f(t) = B^* p(t)$, where p solves the backward adjoint equation.

$$\begin{cases} p'(t) + A^* p(t) = 0, & t \in [0,T] \\ p(T) = p_f, \end{cases} \tag{II.2.1}$$

As already mentioned, the key aspect of our approach hinges on the assertion (II.1.1), where J denotes the so-called *dual* functional, defined by

$$\forall p_f \in \mathbb{R}^n, \qquad J(p_f) := \int_0^T \sigma_{\mathcal{U}}(L_T^* p_f(t))\ \mathrm{d}t + \sigma_{\mathcal{Y}_f}(-p_f) + \langle y_0, e^{TA^*} p_f \rangle. \tag{II.2.2}$$

Remark II.1. When $\mathcal{U}$ is convex, the functional J can be understood as a *dual functional* associated to a primal problem, in the sense of Fenchel-Rockafellar. More details are provided in Appendix I.2. This interpretation leads us to consider useful algorithms that perform a descent over J in order to find dual certificates, as explained in Section II.4.

The following result describes the crucial argument underpinning our method, which is illustrated by Figure II.1.

Proposition II.2. Assume that (H) holds. Then, there exists $p_f \in \mathbb{R}^n$ such that $J(p_f) < 0$ if and only if $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time T.

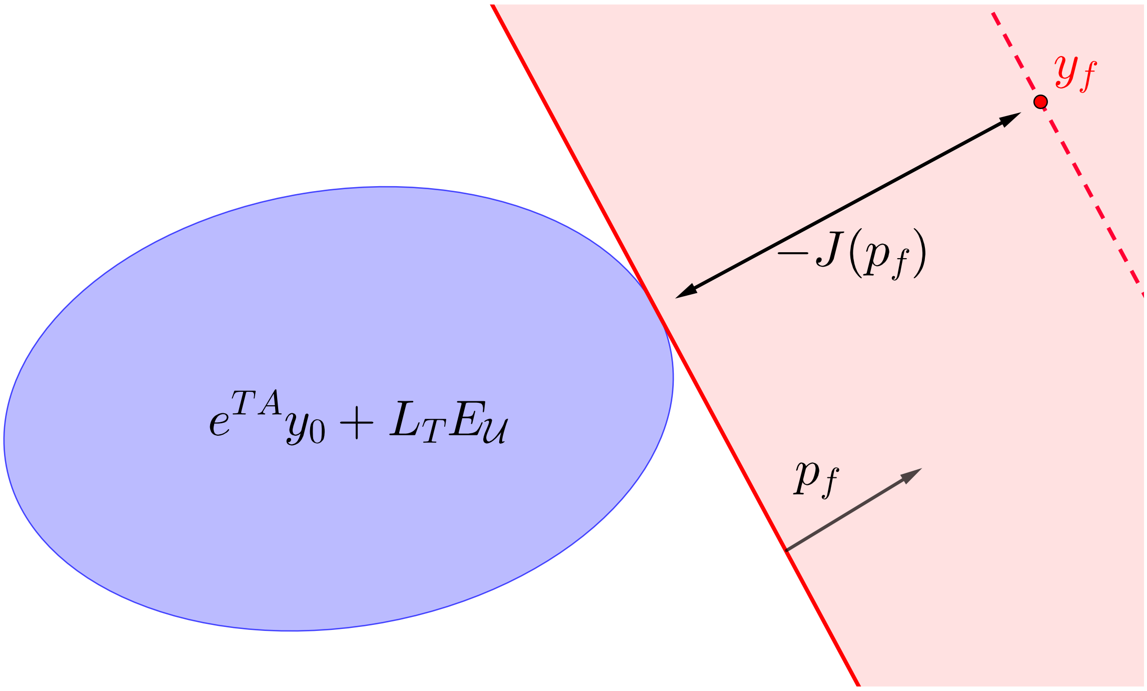


Figure II.1: Reachable set $e^{TA}y_0 + L_T E_\mathcal{U}$, hyperplane associated to the dual certificate p_f, and corresponding scalar $J(p_f)$ given by (II.2.4), for a singleton $\mathcal{Y}_f = \{y_f\}$.

Proof. Let $E_\mathcal{U} := \{u \in E,\ u(t) \in \mathcal{U} \text{ for a.e. } t \in (0,T)\}$. With this notation in place, $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time T if and only if the set $(\mathcal{Y}_f - e^{TA}y_0) \cap L_T E_\mathcal{U}$ is empty, where $\mathcal{Y}_f - e^{TA}y_0 = \{y - e^{TA}y_0, y \in Y_f\}$. Using the basic relation $\sigma_{C-\{y\}}(z) = \sigma_C(z) - \langle y, z\rangle$, we have

$$\sigma_{\mathcal{Y}_f - e^{TA}y_0}(-p_f) = \sigma_{\mathcal{Y}_f}(-p_f) + \langle e^{TA}y_0, p_f\rangle = \sigma_{\mathcal{Y}_f}(-p_f) + \langle y_0, e^{TA^*}p_f\rangle$$

As a result, the function J defined in (II.2.2) rewrites

$$J(p_f) = \int_0^T \sigma_\mathcal{U}(L_T^* p_f(t))\ \mathrm{d}t + \sigma_{\mathcal{Y}_f - e^{TA}y_0}(-p_f) = \sigma_{E_\mathcal{U}}(L_T^* p_f) + \sigma_{\mathcal{Y}_f - e^{TA}y_0}(-p_f),$$

where the interchange of integration and supremum is justified, see e.g. [RW09, Theorem 14.60].

Now assume that we have found p_f such that $J(p_f) < 0$. Then

$$\sigma_{E_\mathcal{U}}(L_T^* p_f) = \sup_{u \in E_\mathcal{U}} \langle u, L_T^* p_f\rangle = \sup_{u \in E_\mathcal{U}} \langle L_T u, p_f\rangle < -\sigma_{\mathcal{Y}_f - e^{TA}y_0}(-p_f) = \inf_{y_f \in \mathcal{Y}_f} \langle y_f - e^{TA}y_0, p_f\rangle,$$

showing that one cannot find $u \in E_\mathcal{U}$ and $y_f \in \mathcal{Y}_f$ such that $L_T u = y_f - e^{TA}y_0$ and hence that $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time $T > 0$.

Conversely, suppose that $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in time T. Then, since $\mathcal{U}$ is compact, it follows from a Lyapunov argument that the set of reachable states (from 0 in time T), i.e., the set

$L_T E_\mathcal{U}$, is compact and convex (see e.g. [LM86, Theorem 1A, Theorem 3 and Lemma 4A in Section 2.2]). The set $\mathcal{Y}_f - e^{TA}y_0$ is closed and convex.

By assumption, these two sets do not intersect, hence we may strictly separate them: there exists $p_f \in \mathbb{R}^n \setminus \{0\}$ such that

$$\sigma_{E_\mathcal{U}}(L_T^* p_f) = \sup_{w\in L_T E_\mathcal{U}} \langle w, p_f\rangle < \inf_{y_f\in\mathcal{Y}_f} \langle y_f - e^{TA}y_0, p_f\rangle = -\sigma_{\mathcal{Y}_f - e^{TA}y_0}(-p_f)$$

which amounts to $J(p_f) < 0$. □

Remark II.3. By positive 1-homogeneity of support functions, J is also positively 1-homogeneous, meaning that $J(\lambda p_f) = \lambda J(p_f)$ for all $\lambda \geq 0$, $p_f \in \mathbb{R}^n$. In particular, if there exists p_f such that $J(p_f) < 0$, then $\inf_{p_f\in\mathbb{R}^n} J(p_f) = -\infty$.

Remark II.4. We could also consider proving that $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from a full set of initial states $\mathcal{Y}_0 \subset \mathbb{R}^n$ in time T, in which case, defining

$$\begin{aligned} \forall\, p_f \in \mathbb{R}^n, \qquad J(p_f) &= \int_0^T \sigma_\mathcal{U}(L_T^* p_f(t))\ \mathrm{d}t + \sigma_{\mathcal{Y}_f}(-p_f) + \sigma_{\mathcal{Y}_0}(e^{TA^*}p_f) \\ &= \int_0^T \sigma_\mathcal{U}(L_T^* p_f(t))\ \mathrm{d}t + \sigma_{\mathcal{Y}_f - e^{TA}\mathcal{Y}_0}(-p_f), \end{aligned}$$

the result of Proposition II.2 holds as it is under the assumption that the set $\mathcal{Y}_f - e^{TA}\mathcal{Y}_0$ is closed and convex; this is the case for instance if $\mathcal{Y}_f$ is closed and convex, and $\mathcal{Y}_0$ is convex and compact.

Remark II.5. The above proposition gives a necessary and sufficient condition for non-reachability. It is worth pointing out that, without any assumptions on the sets $\mathcal{U}$, $\mathcal{Y}_0$, $\mathcal{Y}_f$, the above criterion remains a sufficient condition for non-reachability, as it yields a strict separating hyperplane between $\mathcal{Y}_f$ and $e^{TA}\mathcal{Y}_0 + L_T E_\mathcal{U}$. In that case however, situations where these sets are disjoint but not separable by a hyperplane (typically if $\mathcal{Y}_f$ is not convex) are then undetectable by our approach.

Remark II.6. As mentioned in the introduction, the above can be linked (at least formally) to the Hamilton-Jacobi characterisation of some reachable sets [MBT05; CT18]. Indeed, formally, in optimal control problems, the value function is the solution to a Hamilton Jacobi type equation. Now, for our control problem, the value function is written as

$$S(y_f) := \begin{cases} 0 & \text{if } y_f \text{ is reachable,} \\ +\infty & \text{otherwise,} \end{cases}$$

so we see that the non-reachable set is characterised as the strict zero superlevel set $\{y,\ S(y) > 0\}$ of S. Note that S is a very singular function, and its numerical computation is not tractable, whereas a geometrical approach using support functions leads to a convex function on which a descent algorithm is then implemented, which is much more amenable and prone to numerical certification.

II.2.2 Unsafe sets and minimal times

As previously mentioned, we assume throughout that we know an explicit formula for both functions $\sigma_\mathcal{U}$ and $\sigma_{\mathcal{Y}_f}$, which will be the case in the range of examples we will provide. For instance, for $\mathcal{U}$ defined by the most standard box constraints $\ell_i \leq u_i \leq L_i$ for $i \in \{1, \dots, m\}$, one has with $\ell = (\ell_i)$, $L = (L_i)$ the explicit formula

$$\forall\, u \in \mathbb{R}^m, \quad \sigma_\mathcal{U}(u) = \langle L, u_+\rangle + \langle \ell, u_-\rangle, \tag{II.2.3}$$

where $u_+ = \max(u,0)$ and $u_- = \min(u,0)$ refer to the (componentwise) positive and negative parts of u respectively, and multiplications are to be understood componentwise.

Let us now discuss expressions for the functional (II.2.2) for some specific, yet natural, choices of sets $\mathcal{Y}_f$.

Chosen unsafe sets $\mathcal{Y}_f$. Most of our examples in this article will be based on, but not limited to, the singleton case $\mathcal{Y}_f = \{y_f\}$. Below, we compute the corresponding functional and explain how one then infers results for a closed ball around y_f, *i.e.*, $\mathcal{Y}_f = \overline{B}(y_f, \varepsilon)$, and even a full half-space associated with y_f. Section II.4.3 features a more involved (unbounded) example where $\mathcal{Y}_f$ is a cylinder in $\mathbb{R}^4$, pertaining to the space rendezvous problem.

Singleton. In the case $\mathcal{Y}_f = \{y_f\}$, one computes $\sigma_{\mathcal{Y}_f}(-p_f) = -\langle y_f, p_f\rangle$, which leads to the functional

$$J(p_f) = \int_0^T \sigma_{\mathcal{U}}(L_T^* p_f(t))\,\mathrm{d}t - \langle y_f, p_f\rangle + \langle y_0, e^{TA^*} p_f\rangle. \tag{II.2.4}$$

Ball. In the case of a ball $\mathcal{Y}_f = \overline{B}(y_f, \varepsilon)$ (which recovers the above case with $\varepsilon = 0$), we find

$$\sigma_{\mathcal{Y}_f}(-p_f) = -\langle y_f, p_f\rangle + \varepsilon\|p_f\|,$$

hence we uncover the same functional up to the additional term $\varepsilon\|p_f\|$.

In practice, this has the following implication: given y_f, assume that we have found p_f such that $J(p_f) < 0$ with J given by (II.2.4). Then $\overline{B}(y_f, \varepsilon)$ is not $\mathcal{U}$-reachable from y_0 in time $T > 0$ for any $\varepsilon < -J(\frac{p_f}{\|p_f\|})$. Hence, once a target y_f is fixed, we will only be concerned with the functional J given by (II.2.4). If p_f is found such that $J(p_f) < 0$, we thus obtain a full ball around y_f that is not $\mathcal{U}$-reachable from y_0 in time $T > 0$.

Half-space. We now show how an unreachable half-space can be constructed from any target y_f. For the sake of this remark, when considering the associated functional (II.2.4), we highlight the dependence of J on the target y_f, by writing $J(p_f; y_f)$ instead of just $J(p_f)$.

Now, assume that $\alpha := J(p_f; y_f)$ has been computed for a given $p_f \in \mathbb{R}^n$. For any $\tilde{y_f} \in \mathbb{R}^n$, we have the relation

$$J(p_f; \tilde{y_f}) = J(p_f; y_f) + \langle y_f - \tilde{y_f}, p_f\rangle.$$

Hence, Proposition II.2 shows that, independently of the sign of α, any vector in the half-space

$$\{\tilde{y_f} \in \mathbb{R}^n,\ \langle \tilde{y_f} - y_f, p_f\rangle > \alpha\},$$

is not $\mathcal{U}$-reachable from y_0 in time T. In other words, calculating $J(p_f; y_f)$ for any p_f immediately provides a full half-space that is not $\mathcal{U}$-reachable from y_0 in time T.

Minimal times. It is interesting to notice that, still in the case where $\mathcal{Y}_f = \{y_f\}$ and assuming we have either $y_0 = 0$ or $y_f = 0$, we can also derive a lower bound on the minimal reachability time. We will exploit this result in obtaining (lower) estimates for minimal times in the case of two control systems in Section II.4.

Proposition II.7. Assume that $\mathcal{U} \cap \ker(B) \neq \emptyset$, and suppose either $y_0 = 0$ or $y_f = 0$. If y_f is not $\mathcal{U}$-reachable from y_0 in time T, then it is not reachable for any $\tilde{T} \leq T$ either. Consequently, denoting

$$T^\star(y_0, y_f, \mathcal{U}) = \inf\{T > 0,\ y_f \text{ is } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T\} \in [0 + \infty],$$

we have $T^\star(y_0, y_f, \mathcal{U}) \geq T$.

Proof. This proposition is standard and its proof is elementary. Let us provide the main argument in the case where $y_0 = 0$ for the sake of completeness. Assume that y_f is $\mathcal{U}$-reachable from 0 in time $\tilde{T}$ by a control $\tilde{u}$. Let $T > \tilde{T}$. Let $v \in \mathcal{U} \cap \ker B$. Then, the control u defined by $u(t) = v$ for $t \in (0, T - \tilde{T})$ and $u(t) = \tilde{u}(t - T + \tilde{T})$ steers the system from 0 to y_f in time T and satisfies the constraint, hence the conclusion. The end of the proof is straightforward. □

Remark II.8. In the context of proving that a given unsafe set $\mathcal{Y}_f$ is not reachable for all times $t \in [0,T]$, let us note the following equivalence: $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in all times $t \in [0,T]$ if and only if for all $t \in [0,T]$, $\mathcal{Y}_f - e^{tA}y_0$ is not $\mathcal{U}$-reachable from 0 in time t. This leads to a sufficient condition: if $\mathcal{Y}_f - \cup_{t\in[0,T]}\{e^{tA}y_0\}$ is not $\mathcal{U}$-reachable from 0 in time T, then $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from y_0 in all times $t \in [0,T]$.

II.3 Discretisation and error estimates

This section presents several discretisations and corresponding error estimates for the dual functional (II.2.2). Error estimates are given using standard Hermitian norms (over $\mathbb{C}^n$ and $\mathbb{C}^m$), always denoted by $\|\cdot\|$. The same notation $\|\cdot\|$ will also be used for the corresponding operator norms, that of matrices in $\mathbb{C}^{n\times n}$, $\mathbb{C}^{m\times n}$ and $\mathbb{C}^{n\times m}$.

As discussed in the introduction, we make the reasonable assumption that we have access to an explicit formula for $\sigma_{\mathcal{U}}$ (and $\sigma_{\mathcal{Y}_f}$). Also recall that $\mathcal{U}$ is compact, and M denotes a positive constant such that $\|v\| \le M$ for all $v \in \mathcal{U}$. In particular, we can easily prove that $|\sigma_{\mathcal{U}}(x) - \sigma_{\mathcal{U}}(y)| \le |\sigma_{\mathcal{U}}(x-y)| \le M\|x-y\|$, which implies that $\sigma_{\mathcal{U}}$ is M-Lipschitz.

II.3.1 Partial discretisation for a known adjoint exponential

To evaluate the dual functional (II.2.2) at a given point p_f, one must compute a time integral, and solve the backward equation (II.2.1). Given that $\sigma_{\mathcal{U}}$ will generally not be better behaved than Lipschitz, we will stick to time-discretisation schemes that are of order 1, whether for computing integrals or for integrating ODEs.

Even when one has access to an explicit solution for the backward equation (II.2.1), the integral will seldom be computable (or at the cost of cumbersome computations). This is why we first consider the case of discretising the integral but not the backward equation (II.2.1).

We define

$$N_0 \in \mathbb{N}^*, \quad \Delta t = \frac{T}{N_0}, \quad t_k = k\Delta t \text{ for } k \in \{0,\dots,N_0\}.$$

For a fixed $p_f \in \mathbb{R}^n$, we let $t \mapsto p(t)$ be the solution to (II.2.1), i.e., $p(t) = e^{(T-t)A^*}p_f$, and consider $J_{\mathrm{d},1}$, the first discretised version of J given by

$$J_{\mathrm{d},1}(p_f) := \Delta t \sum_{k=1}^{N_0} \sigma_{\mathcal{U}}(B^* p(t_k)) + \sigma_{\mathcal{Y}_f}(-p_f) + \langle y_0, p(0)\rangle. \tag{II.3.1}$$

Proposition II.9. For a given $p_f \in \mathbb{R}^n$, it holds that

$$|J(p_f) - J_{\mathrm{d},1}(p_f)| \le \frac{1}{2}\Delta t\ MT\|B\|\Big(\sup_{t\in[0,T]} \|e^{tA^*}\|\Big)\|A^*p_f\|.$$

Proof. Recall that $\sigma_{\mathcal{U}}$ is M-Lipschitz continuous; therefore, we have for all $s,t \in [0,T]$

$$|\sigma_{\mathcal{U}}(L_T^* p_f(s)) - \sigma_{\mathcal{U}}(L_T^* p_f(t))| \le M\|B^*p(s) - B^*p(t)\| \le M\|B\|\|p(t) - p(s)\|.$$

We can now establish the bound

$$|\sigma_{\mathcal{U}}(L_T^* p_f(s)) - \sigma_{\mathcal{U}}(L_T^* p_f(t))| \le M\|B\| \sup_{t\in[0,T]} \|A^*p(t)\|\,|t-s|.$$

We have proved that $t \mapsto \sigma_{\mathcal{U}}(L_T^* p_f(t))$ is Lipschitz continuous. Recalling the standard estimate

$$\left|\int_0^T f(t)\ \mathrm{d}t - \Delta t \sum_{k=1}^{N_0} f(t_k)\right| \le \frac{1}{2} KT\,\Delta t$$

for a K-Lipschitz function $f : [0,T] \to \mathbb{R}$, we end up with

$$\left| \int_0^T \sigma_{\mathcal{U}}(B^* p(t))\,\mathrm{d}t - \Delta t \sum_{k=1}^{N_0} \sigma_{\mathcal{U}}(B^* p(t_k)) \right| \leq \frac{1}{2} \Delta t M T \|B\| \sup_{t\in[0,T]} \|A^* p(t)\|,$$

thus, the previously announced estimate readily follows, using the definition of $p(t)$:

$$\begin{aligned} \sup_{t\in[0,T]} \|A^* p(t)\| &= \sup_{t\in[0,T]} \|A^* e^{(T-t)A^*} p_f\| = \sup_{t\in[0,T]} \|A^* e^{tA^*} p_f\| \\ &= \sup_{t\in[0,T]} \|e^{tA^*} A^* p_f\| \leq \sup_{t\in[0,T]} \|e^{tA^*}\| \|A^* p_f\|. \end{aligned}$$

□

Jordan-Chevalley decomposition. Even if one knows the matrix exponentials $t \mapsto e^{tA^*}$ (or equivalently the matrix exponentials $t \mapsto e^{tA}$), it is still necessary to provide an upper bound for $\sup_{t\in[0,T]} \|\, e^{tA^*}\| = \sup_{t\in[0,T]} \|e^{tA}\|$ to make the bound in Proposition II.9 useful.

Assume that we have access to the Jordan-Chevalley decomposition of A in the following sense: we have $A = D + N$ where D is diagonalisable, N is nilpotent with index ℓ, the two matrices D and N commute. Then, of course, e^{tA} is obtained by

$$\forall\, t \in \mathbb{R}, \qquad e^{tA} = e^{tD} \sum_{k=0}^{\ell-1} \frac{N^k}{k!} t^k = e^{tD} Q_\ell(tN), \tag{II.3.2}$$

where Q_ℓ is the polynomial $x \mapsto \sum_{k=0}^{\ell-1} \frac{x^k}{k!}$. Assume further that we have access to the transition matrix P that diagonalises D, i.e., $\mathrm{diag}(\Lambda) = P^{-1} D P$ where $\Lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{C}^n$ stores the eigenvalues of A.

Consequently, we have

$$e^{tA} = P e^{t\Lambda} P^{-1} Q_\ell(tN),$$

which leads to the estimate

$$\sup_{t\in[0,T]} \|e^{tA}\| \leq \kappa(P) e^{\mu T} Q_\ell(\|N\| T),$$

where $\mu := \max(\{\mathrm{Re}(\lambda_i),\ i \in \{0, \dots, n\}\})$ is the spectral abscissa of A, and $\kappa(P) = \|P\| \|P^{-1}\|$ stands for the condition number of the transition matrix P.

From these estimates, we derive the error formula below, in the case where the Jordan-Chevalley decomposition is known.

Corollary II.10. Let us assume that we know the explicit Jordan-Chevalley decomposition of A, in the form $A = D + N$. Then for a given $p_f \in \mathbb{R}^n$, there holds

$$|J(p_f) - J_{\mathrm{d},1}(p_f)| \leq \frac{1}{2} \Delta t\ M T \|B\| \|A^* p_f\| \kappa(P) e^{\mu T} Q_\ell(\|N\| T).$$

II.3.2 Full discretisation

We now address the scenario where the adjoint exponential $t \mapsto e^{tA^*}$ is unknown, necessitating the discretisation of the backward equation (II.2.1) as well. Assume that a discretisation scheme has been applied that produces $p_k \in \mathbb{R}^n$ for $k \in \{0, \dots, N_0\}$.

In the next subsection, we will specialise to the Euler implicit scheme for the class of negative semi-definite matrices.

The fully discretised version of J then reads

$$J_{\mathrm{d},2}(p_f) := \Delta t \sum_{k=1}^{N_0} \sigma_{\mathcal{U}}(B^* p_k) + \sigma_{\mathcal{Y}_f}(-p_f) + \langle y_0, p_0 \rangle. \tag{II.3.3}$$

Proposition II.11. For a given $p_f \in \mathbb{R}^n$ and vectors $p_k \in \mathbb{R}^n$, $k \in \{0, \dots, N_0\}$, there holds

$$|J(p_f) - J_{\mathrm{d},2}(p_f)| \le \Delta t\ M\|B\| \left(\frac{1}{2}T\|A^*p_f\| \sup_{t\in[0,T]} \|e^{tA^*}\| + \sum_{k=1}^{N_0} \|p(t_k) - p_k\|\right) + \|y_0\|\|p(0) - p_0\|.$$

The proof is straightforward and left to the reader, as it primarily involves providing an estimate for $|J_{\mathrm{d},1}(p_f) - J_{\mathrm{d},2}(p_f)|$ and combining it with the estimate given from Proposition II.9.

We now explore the application of the simplest possible scheme, which is the Euler explicit scheme:

$$\begin{cases} p_{N_0} = p_f \\ p_k = (\mathrm{Id} + \Delta t A^*)p_{k+1} \quad \forall\, k \in \{0, \dots, N_0 - 1\}. \end{cases} \tag{II.3.4}$$

Note that the Euler implicit scheme could also be employed and would yield similar results. It is then standard (see e.g. [QSS06, Section 11.3.2]) that

$$\forall\, k \in \{0, \dots, N_0\}, \quad \|p(t_k) - p_k\| \le \frac{1}{2}\Delta t\ (T - t_k)\Big(\sup_{t\in[t_k,T]} \|p''(t)\|\Big)e^{\|A\|T}.$$

Given that $p''(t) = e^{(T-t)A^*}(A^*)^2 p_f$, this leads to the estimate

$$\begin{aligned} \forall\, k \in \{0, \dots, N_0\}, \quad \|p(t_k) - p_k\| &\le \frac{1}{2}\Delta t\ (T - t_k)\Big(\sup_{t\in[t_k,T]} \|e^{tA^*}\|\Big)e^{\|A\|T}\|(A^*)^2 p_f\| \\ &\le \frac{1}{2}\Delta t\ (T - t_k)e^{2\|A\|T}\|(A^*)^2 p_f\| \end{aligned}$$

We acknowledge that constants appearing in the above might slightly be improved.

All in all, we thus find the following global estimate.

Proposition II.12. For a given $p_f \in \mathbb{R}^n$ and vectors $p_k \in \mathbb{R}^n$, $k \in \{0, \dots, N_0\}$ defined according to the Euler explicit scheme (II.3.4), it holds that

$$|J(p_f) - J_{\mathrm{d},2}(p_f)| \le \frac{1}{2}\Delta t\, T\left[M\|B\|\left(e^{\|A\|T}\|A^*p_f\| + \frac{1}{2}Te^{2\|A\|T}\|(A^*)^2 p_f\|\right) + \|y_0\|e^{2\|A\|T}\|(A^*)^2 p_f\|\right].$$

Proof. The only step that requires detailed explanation is the estimation of the sum of the errors $\|p(t_k) - p_k\|$, obtained by writing

$$\sum_{k=1}^{N_0} \|p(t_k) - p_k\| \le \frac{1}{2}\Delta t\, e^{2\|A\|T}\|(A^*)^2 p_f\| \sum_{k=1}^{N_0}(T - t_k) = \frac{1}{2}\Delta t\, e^{2\|A\|T}\|(A^*)^2 p_f\| \frac{T}{N_0}\sum_{k=1}^{N_0}(N_0 - k).$$

The sum $\sum_{k=1}^{N_0}(N_0 - k)$ equals $\frac{(N_0-1)N_0}{2}$; therefore

$$\sum_{k=1}^{N_0} \|p(t_k) - p_k\| = \frac{1}{4}\Delta t\, e^{2\|A\|T}\|(A^*)^2 p_f\| T(N_0 - 1) \le \frac{1}{4}T^2 e^{2\|A\|T}\|(A^*)^2 p_f\|.$$

□

This estimate has one major drawback: it diverges exponentially fast as a function of T, making the investigation of non-$\mathcal{U}$-reachability challenging, even for moderate times $T > 0$, especially if the matrix norm $\|A\|$ is large.

II.3.3 Full discretisation for a symmetric negative semidefinite matrix

The purpose of this subsection is to exhibit a class of matrices, that of symmetric negative semidefinite matrices, for which refined estimates can be derived without the errors exponentially diverging as a function of time T.

Even though such matrices are diagonalisable, computing their exponential can become intractable for large sizes, so that one needs to resort to discretisation for the backward equation (II.2.1). The implicit Euler scheme below is well suited to that situation:

$$\begin{cases} p_{N_0} = p_f \\ (\mathrm{Id} - \Delta t A^*) p_k = p_{k+1} \quad \forall\, k \in \{0, \dots, N_0 - 1\}. \end{cases} \tag{II.3.5}$$

It always makes sense provided Δt is small enough, and in the case where the matrix A is a negative semidefinite symmetric matrix, the Euler implicit scheme is well-defined whatever the value of $\Delta t > 0$.

Assuming we are given a symmetric positive semidefinite matrix C, diagonalised in the form $C = PDP^{-1}$, with D diagonal and P an orthogonal transition matrix, we may define $\varphi(C)$ for any function $\varphi : [0, +\infty) \to \mathbb{R}$ by $\varphi(C) = P\varphi(D)P^{-1}$ with componentwise application of φ on the diagonal. This definition obviously agrees with the usual matrix exponential and rational fractions whose poles avoid $[0, +\infty)$.* Using that $\kappa(P) = 1$, one has for all such functions

$$\|\varphi(C)\| = \|\varphi(D)\| \le \sup_{x \ge 0} |\varphi(x)|, \tag{II.3.6}$$

Proposition II.13. Assume that A is a negative semidefinite symmetric matrix, and let $p_f \in \mathbb{R}^n$. Then the error between the solution to the backward ODE (II.2.1) and its implicit Euler discretisation (II.3.5) satisfies

$$\forall\, k \in \{0, \dots, N_0\}, \quad \|p(t_k) - p_k\| \le \frac{1}{2} \Delta t\, \|A^* p_f\|. \tag{II.3.7}$$

Proof. By definition, for all $k \in \{0, \dots, N_0\}$, we have

$$p(t_k) - p_k = \left[e^{(T - t_k)A^*} - (\mathrm{Id} - \Delta t A^*)^{-(N_0 - k)} \right] p_f.$$

Hence me may write

$$p(t_k) - p_k = -\Delta t\, \varphi_{N_0 - k}(-\Delta t A^*) A^* p_f,$$

where for $k \in \mathbb{N}^*$, the function φ_k is defined for $x > 0$ by

$$\varphi_k(x) := \frac{e^{-kx} - (1+x)^{-k}}{x},$$

extended by continuity at $x = 0$ by $\varphi_k(0) := 0$.

Estimating, we find

$$\|p(t_k) - p_k\| \le \Delta t\, \|\varphi_{N_0 - k}(-\Delta t A^*)\|\, \|A^* p_f\| \le \Delta t \sup_{x \ge 0} |\varphi_{N_0 - k}(x)| \|A^* p_f\|$$

Let us conclude by proving that $\sup_{x \ge 0} |\varphi_k(x)| \le \frac{1}{2}$ for all $k \ge 1$. First, a routine study shows that the function $x \mapsto e^{-x}(1+x) - 1 + \frac{1}{2}x^2$ is nonnegative for all $x \ge 0$, so that

$$|\varphi_1(x)| = \frac{1}{x} \left[\frac{1}{1+x} - e^{-x} \right] \le \frac{1}{2} x, \tag{II.3.8}$$

which combined with the basic estimate $|\varphi_1(x)| \le \frac{1}{x} \frac{1}{1+x}$ for $x > 0$ yields $|\varphi_1(x)| \le \frac{1}{2}$ by considering the two cases $x \le 1$ and $x > 1$. Now for $k \ge 2$, and $x > 0$, we write

$$\begin{aligned} |\varphi_k(x)| = \frac{1}{x} \left[\frac{1}{1+x} \right] \sum_{j=0}^{k-1} e^{-jx} \left(\frac{1}{1+x} \right)^{k-j-1} &= |\varphi_1(x)| \sum_{j=0}^{k-1} e^{-jx} \left(\frac{1}{1+x} \right)^{k-j-1} \\ &\le |\varphi_1(x)| \frac{k}{(1+x)^{k-1}}. \end{aligned}$$

[*] There are of course much more general definitions for functions of matrices [Hig08], but in the present setting this definition will suffice.

thanks to the bound $e^{-x} \leq \frac{1}{1+x}$. Let us focus on the case $k = 2$. If $x \leq 1$, we have $|\varphi_2(x)| \leq \frac{1}{2}x\frac{2}{1+x} \leq \frac{1}{2}$, and for $x > 1$, $|\varphi_2(x)| \leq \frac{1}{x(1+x)}\frac{2}{1+x} \leq \frac{1}{2}$, hence the result for $k = 2$.

Now for any $k \geq 3$, using the estimate (II.3.8), we obtain the inequality

$$|\varphi_k(x)| \leq \frac{kx}{2(1+x)^{k-1}}$$

The right-hand side is maximised at $x = \frac{1}{k-2}$, hence

$$|\varphi_k(x)| \leq \frac{k}{2(k-2)}\left(\frac{k-2}{k-1}\right)^{k-1} = \frac{1}{2}\frac{k(k-2)}{(k-1)^2}\left(\frac{k-2}{k-1}\right)^{k-3} \leq \frac{1}{2}.$$

□

This entails the following compact estimate for the dual functional.

Proposition II.14. Assume that A is a symmetric negative semidefinite matrix. For a given $p_f \in \mathbb{R}^n$ and vectors $p_k \in \mathbb{R}^n$, $k \in \{0, \ldots, N_0\}$ defined according to the Euler implicit scheme (II.3.5), there holds

$$|J(p_f) - J_{\mathrm{d},2}(p_f)| \leq \Delta t \, \|A^* p_f\| \left(TM\|B\| + \frac{1}{2}\|y_0\|\right). \tag{II.3.9}$$

Proof. We simply build upon the general estimate of Proposition II.11. First, since $-A^*$ is a symmetric positive semidefinite matrix, (II.3.6) provides

$$\|e^{tA^*}\| \leq \sup_{x \geq 0} |e^{-tx}| = 1$$

for all $t \geq 0$, and the previous estimate from Proposition II.13 for the Euler implicit scheme shows that

$$\sum_{k=1}^{N_0} \|p(t_k) - p_k\| \leq \frac{1}{2}\Delta t\|A^* p_f\|N_0 = \frac{1}{2}T\|A^* p_f\|.$$

Remark II.15. We note that similar estimates, not exponentially diverging with T, could also be derived for the broader class of dissipative matrices (i.e., matrices A satisfying $\langle Ax, x\rangle \leq 0$ for all $x \in \mathbb{R}^n$).

□

II.4 Numerical approach and examples

In this section, we will illustrate the potential of the approach described in the previous section to study the (non)-reachability of certain targets, in a variety of examples. We present three main example families, respectively related to the following:

- The control of a streetcar that we wish to control in order to reach a final state in minimal time. This is a well-known toy problem in optimal control theory. We use it to validate our results since the reachable set and minimal times (from $(0,0)^T$) have known explicit formulae.

- The spatial rendezvous problem. We aim at reaching (or avoiding) a given target, corresponding to a space station, for instance the ISS, in a referential centred in the initial position of the spacecraft. We use a dynamic space mechanics model and provide certified lower-bounds on the minimal time needed to reach the target. We then develop a method to prove that the spacecraft cannot collide with a motionless obstacle – e.g. an asteroid – within a predetermined time interval.

- A more academic setting, based on randomly generated negative semi-definite (Jacobi) matrices A (of possibly large dimension). This is designed to investigate cases where computing exponentials becomes out of reach, as well as to explore the effect that increasing the dimension has on our technique.

Most cases feature a set of the form $\mathcal{Y}_f = \{y_f\}$, hence the function of interest is (II.2.4). As explained in Subsection II.2, the use of the corresponding functional also allows us to certify that balls around y_f or even half-spaces cannot be reached. The types of constraint sets $\mathcal{U}$ also vary across examples.

II.4.1 Numerical approach and methodology

In order to numerically verify the non-$\mathcal{U}$-reachability of a given target y_f from y_0 in time T, one must proceed through the following three steps:

1. First, one must compute a discretisation J_d of the functional J, for example $J_{\mathrm{d},1}$ or $J_{\mathrm{d},2}$, with the associated bounds on discretisation errors

2. Then, one must minimise said discretisation in order to find an element p_f such that $J_\mathrm{d}(p_f) < 0$.

3. Finally, one must compute $e(p_f)$ such that $J_\mathrm{d}(p_f) - e(p_f) \leq J(p_f) \leq J_\mathrm{d}(p_f) + e(p_f)$. This is done here using the INTLAB toolbox [Rum99], which, using interval arithmetic, takes into account the rounding errors and added discretisation errors. This leads to the verification that indeed, $J(p_f) \leq J_\mathrm{d}(p_f) + e(p_f) < 0$. If that is not the case, either y_f is reachable, or a finer discretisation or minimisation is required to prove its non-reachability.

Since INTLAB allows for most typical computation techniques, the second and third steps could be joined. However, interval arithmetic is computationally expensive, hence we first minimise the discretised functional J_d to find p_f such that $J_\mathrm{d}(p_f) < -\eta$, where η is the typical size of errors $e(p_f)$ (on the ball $\|p_f\| = 1$), and then verify that p_f is indeed a certificate of non-$\mathcal{U}$-reachability for y_f. Since this stopping condition will never be satisfied if the target set $\mathcal{Y}_f$ is in fact $\mathcal{U}$-reachable, one might consider adding another condition based on how small an improvement is made from one step to another. As the functional J_d does not admit a minimiser (see Remark II.3) in the non-reachable case, we use the stopping condition

$$\left\| \frac{p_{k+1}}{\|p_{k+1}\|} - \frac{p_k}{\|p_k\|} \right\| \leq \delta, \tag{II.4.1}$$

where δ is a small tolerance.

Carrying out a descent algorithm on J_d can be tackled by means of many optimisation techniques. For the following examples, we take advantage of the dual nature (see Appendix I.2, assuming $\mathcal{U}$ is convex. This allows us to use the Chambolle-Pock primal-dual algorithm [CP11]. It has the drawback of requiring a closed-form expression of two proximal operators associated with the functionals F^* and G, as defined in Appendix I.2. In general, if $\sigma_\mathcal{U}$ and $\sigma_{\mathcal{Y}_f}$ have closed-form formulae, so do those proximal operators.

II.4.2 The streetcar

Control problem. The following example is completely standard in optimal control theory. It can be found, for example, in [LM86, Chapter 1] and is concerned with the optimal control of the acceleration of a streetcar on a straight axis.

We will use this example to both illustrate and validate our approach, since the reachable set and minimal times are known explicitly, see Appendix II.5.

We consider a streetcar moving on a graduated rectilinear axis. The initial position-velocity pair of the streetcar is assumed to be $(0, 0)^T$ and the objective is to steer the system to some

$y_f \in \mathbb{R}^2$ in minimal time. The control system reads

$$\begin{cases} y_1'(t) = y_2(t), \\ y_2'(t) = u(t), \end{cases} \tag{II.4.2}$$

which corresponds to the matrices

$$A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B := \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{II.4.3}$$

For a fixed $M > 0$, the chosen constraint is given by

$$\mathcal{U} := \{u \in \mathbb{R},\ |u| \leq M\}.$$

Resolution method. First, we compute the support function

$$\forall\, u \in \mathbb{R}, \quad \sigma_{\mathcal{U}}(u) = M|u|,$$

which is a particular case of (II.2.3).

Here, we use the functional $J_{\mathrm{d},1}$ and the estimate given by Corollary II.10. Given how simple $\sigma_{\mathcal{U}}$ and the control system are, we acknowledge that one could actually compute the functional J itself and only have to deal with round-off errors. We do not pursue this approach since we aim at analysing how prominent the discretisation errors may be.

The Jordan-Chevalley decomposition of A is straightforward in this case, since the matrix A is itself nilpotent, of index $\ell = 2$. In this case, we hence have $\mu = 0$, $\kappa(P) = 1$, $\ell = 2$, $Q_2(x) = 1 + x$, leading to the estimate

$$|J(p_f) - J_{\mathrm{d},1}(p_f)| \leq \frac{1}{2}\Delta t\ MT\|B\|\|A^* p_f\|Q_2(\|A\|T).$$

Results. To highlight the dependence of J with respect to the target y_f, we will temporarily rename $J(p_f)$ to $J(p_f; y_f)$. We give examples of targets $y_f \in \mathbb{R}^2$ that are certified to not be $\mathcal{U}$-reachable below, in the form of a computer-assisted theorem.

Theorem II.16. The following targets are not $\mathcal{U}$-reachable from $(0,0)$ in time $T = 1$, with $M = 1$:

$$y_1 = (0.1, 0.6)^T, \qquad y_2 = (0.5, 1.1)^T, \qquad y_3 = (0.3, 0)^T.$$

Indeed, the dual certificates

$$p_1 = (-0.77, 0.64)^T, \qquad p_2 = (0.29, 0.96)^T, \qquad p_3 = (0.85, -0.53)^T.$$

provide the intervals

$$J(p_1; y_1) \in [-0.0305, -0.0291], \quad J(p_2; y_2) \in [-0.0964, -0.0959], \quad J(p_3; y_3) \in [-0.0282, -0.0268].$$

The targets and dual certificates are plotted in Figure II.2, along with the theoretically known reachable set.

Using the formula provided in Appendix II.5, the minimal times to reach y_1, y_2, and y_3 are computed to be slightly above 1.1656, 1.7480, and 1.0954, which means they are indeed not reachable.

II.4.3 Space rendezvous

Control problem. We here consider the 2-dimensional linearised Hill-Clohessy-Wiltshire equations, as defined in [CF18]. These equations model the motion of a follower spacecraft in the

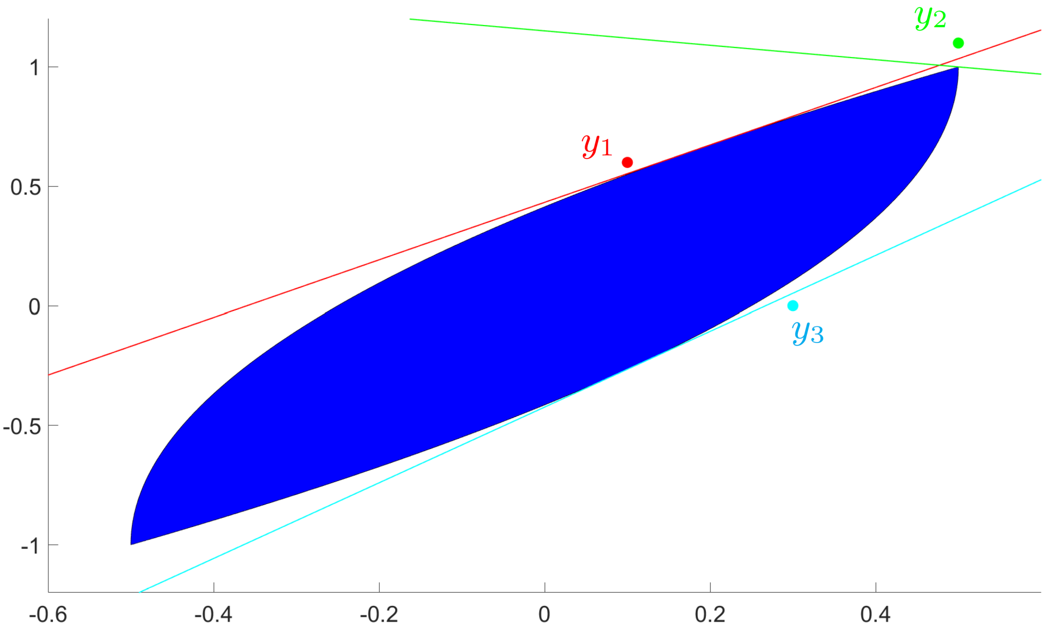


Figure II.2: Non-$\mathcal{U}$-reachability of the targets from Theorem II.16, together with the support hyperplane associated to their respective dual certificates, and the streetcar theoretical reachable set deduced from Appendix II.5.

neighbourhood of a reference spacecraft (at position $y_0 = (0, 0, 0, 0)^T$). The matrices underlying the control problem are

$$A:=\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 2 \\ 0 & 0 & -2 & 0 \end{pmatrix}, \qquad B:=\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{II.4.4}$$

Note that y_1, y_2 are positions and $y_3 = y_1'$, $y_4 = y_3'$ are the corresponding speeds.

We consider the following constraint set for fixed $M_2 > 0$, $M_\infty > 0$:

$$\mathcal{U}:=\{u \in \mathbb{R}^2,\ \|u\|_2 \le M_2,\ \|u\|_\infty \le M_\infty\}, \tag{II.4.5}$$

hence we may take $M:=\min(M_2, \sqrt{2}M_\infty)$.

Let us compute the support function $\sigma_{\mathcal{U}}$ in the case where $M_\infty \le M_2 \le \sqrt{2}M_\infty$, which we will consider hereafter. As illustrated in Figure II.3, the constraint set is the intersection of a disk and a square. Observe that the boundary of $\mathcal{U}$ is the union of flat and circular parts, whose coordinates (x, y) of intersection points form the set

$$\mathcal{P} = \left\{\left(\pm M_\infty, \pm\sqrt{M_2^2 - M_\infty^2}\right)\right\} \bigcup \left\{\left(\pm\sqrt{M_2^2 - M_\infty^2}, \pm M_\infty\right)\right\}.$$

Let us write $\partial\mathcal{U} = \mathcal{F} \cup \mathcal{C}$, where $\mathcal{F}$ (resp. $\mathcal{C}$) denotes the union of all flat (resp. circular) parts of the boundary.

Let us fix $x \in \mathbb{R}^2$. We distinguish between two cases:

- If $(O; x) \cap \partial\mathcal{U} \subset \mathcal{C}$, meaning that $\frac{\|x\|_\infty}{M_\infty} \le \frac{\|x\|_2}{M_2}$, then $M_2\frac{x}{\|x\|_2} \in \mathcal{U}$ and using the Cauchy-Schwarz inequality, we get

$$\sigma_{\mathcal{U}}(x) \le \sup_{y\in\mathcal{U}} \|x\|_2\|y\|_2 = \left\langle x, M_2\frac{x}{\|x\|_2}\right\rangle = M_2\|x\|_2.$$

  We thus infer that $\sigma_{\mathcal{U}}(x) = M_2\|x\|_2$.

- Otherwise, $\sigma_{\mathcal{U}}(x)$ reads as the maximum of a linear (convex) function on a union of flat parts. We easily infer that $\sigma_{\mathcal{U}}(x) = \langle p_x, x\rangle$, where p_x denotes any point of the set $\operatorname{argmin}_{p\in\mathcal{P}} \|p - x\|_2$.

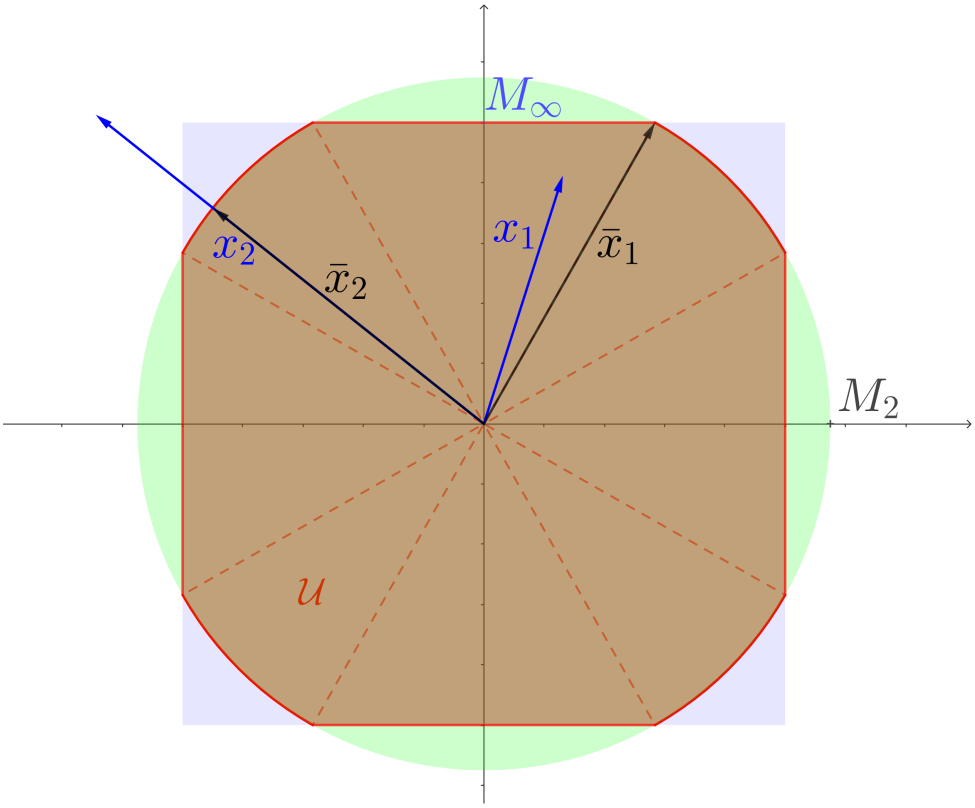


Figure II.3: Construction of the support function for the rendezvous problem. One has in particular $\sigma_{\mathcal{U}}(x_i) = \langle x_i, \bar{x}_i \rangle$, $i = 1, 2$.

Resolution method. The Jordan-Chevalley decomposition of A is given by $A = D + N$ with

$$D{:=}P\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&-i&0\\0&0&0&i\end{pmatrix}P^{-1},\quad N{:=}P\begin{pmatrix}0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{pmatrix}P^{-1},\quad P{:=}\begin{pmatrix}0&-\frac{2}{3}&-1&-1\\1&0&2i&-2i\\0&0&i&-i\\0&1&2&2\end{pmatrix}.$$

Here, we also use the functional $J_{\mathrm{d},1}$ and the estimate given by Corollary II.10. Using the corresponding notations, we have $\mu = 0$, and the index of the nilpotent matrix N is $\ell = 2$. Thus the corresponding estimate reads

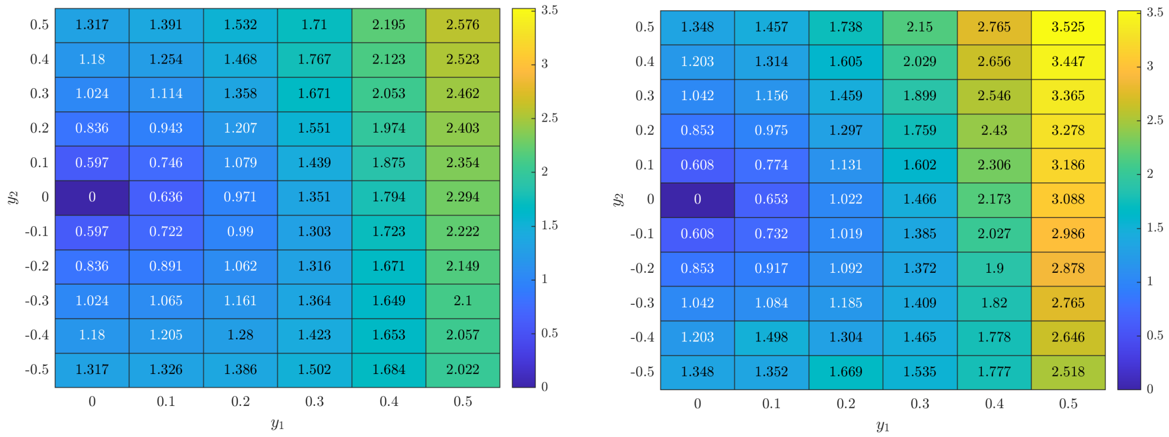
$$|J(p_f) - J_{\mathrm{d},1}(p_f)| \leq \frac{1}{2}\Delta t\ MT\|B\|\|A^* p_f\|\kappa(P)Q_2(\|N\|T),$$

with $Q_2(x) = 1 + x$.

Results. Given a target $y_f \in \mathbb{R}^4$, we can derive a lower-bound on the minimal time needed to steer the system from $y_0 = (0, 0, 0, 0)^T$ to y_f. Proposition II.7 ensures that we may indeed estimate the corresponding minimal time from below, using our approach. To compute this lower bound, we apply a bisection algorithm over the set of positive real numbers, starting from a predefined interval $[t_{\mathrm{inf}}, t_{\mathrm{sup}}]$, and expanding it by multiplying its length by 2 until we cannot prove the non-reachability in time t_{sup}, and we can prove it in time t_{inf}. Then, the standard bisection method applies until the interval is reduced to the desired length.

First, we consider the time-minimal control problem of steering the system from $y_0 = 0$ to some other position at 0 speed, i.e., $y_f = (y_1, y_2, 0, 0)^T$ for various values of $(y_1, y_2) \in \mathbb{R}^2$. Since the control problem is linear and the constraints centrally symmetric (i.e., $\mathcal{U} = -\mathcal{U}$), if y_f is reachable in time $T > 0$, so is $-y_f$. This translates into the identity $J(p_f; y_f) = J(-p_f; -y_f)$, allowing us to focus our computations on the right half-plane.

Using the bounds $M_2 = 1.15$ and $M_\infty = 1$, we obtain the certified lower bounds on the minimal-time shown on Figure II.4(a). For conciseness, we do not provide the corresponding



	0	0.1	0.2	0.3	0.4	0.5
0.5	1.317	1.391	1.532	1.71	2.195	2.576
0.4	1.18	1.254	1.468	1.767	2.123	2.523
0.3	1.024	1.114	1.358	1.671	2.053	2.462
0.2	0.836	0.943	1.207	1.551	1.974	2.403
0.1	0.597	0.746	1.079	1.439	1.875	2.354
0	0	0.636	0.971	1.351	1.794	2.294
-0.1	0.597	0.722	0.99	1.303	1.723	2.222
-0.2	0.836	0.891	1.062	1.316	1.671	2.149
-0.3	1.024	1.065	1.161	1.364	1.649	2.1
-0.4	1.18	1.205	1.28	1.423	1.653	2.057
-0.5	1.317	1.326	1.386	1.502	1.684	2.022

	0	0.1	0.2	0.3	0.4	0.5
0.5	1.348	1.457	1.738	2.15	2.765	3.525
0.4	1.203	1.314	1.605	2.029	2.656	3.447
0.3	1.042	1.156	1.459	1.899	2.546	3.365
0.2	0.853	0.975	1.297	1.759	2.43	3.278
0.1	0.608	0.774	1.131	1.602	2.306	3.186
0	0	0.653	1.022	1.466	2.173	3.088
-0.1	0.608	0.732	1.019	1.385	2.027	2.986
-0.2	0.853	0.917	1.092	1.372	1.9	2.878
-0.3	1.042	1.084	1.185	1.409	1.82	2.765
-0.4	1.203	1.498	1.304	1.465	1.778	2.646
-0.5	1.348	1.352	1.669	1.535	1.777	2.518

Figure II.4: Estimates of the minimal time for reachability of various targets at speed 0 for the spacecraft rendezvous control problem. Certified lower bounds (left panel (a)) versus minimal times outputted by Gekko Optimization Suite [Bea+18] (right panel (b)).

dual certificates. For comparison purposes, the minimal times computed using the Python package Gekko [Bea+18] are presented in Figure II.4(b). Note that Gekko does not control discretisation bounds nor roundoff errors, hence the corresponding estimates are by no means certified.

Computation times. As is common, our certified method comes at the price of increased computation times: each step of the bisection algorithm is rather fast (about 30 seconds on a standard desktop computer), but depending on parameters and how good the initial guess is, the number of iterations of the bisection algorithm may go from 3-4 to 10-15 iterations, whereas Gekko's method computes one approximation of the minimal time in about 10 seconds.

Assuming that Gekko produces reliable estimates, the accuracy of our method seems to decrease the further the target y_f is from y_0, going from about 1.8% to 37%. This can be explained as follows: our computations were made with a fixed number of time steps, namely $N_0 = 20,000$; hence the higher the theoretical minimal time is, the harder it is to establish a tight lower-bound. Increasing N_0 allows for a more precise approximation: for example, for $y_f = (0.5, 0.5, 0, 0)^T$, with $N_0 = 400,000$, the dual certificate $p_f = (0.874, 0.0914, -0.3008, 0.3704)^T$ proves the bound $t_{\min} \geq 3.4$, which is about 3.7% away from Gekko's approximation.

On the other hand, Gekko seems to produce what might be artifacts (points $(0.1, -0.4)^T$ and $(0.2, -0.5)^T$), while our computed certified lower bounds remain smooth.

More complex unsafe set $\mathcal{Y}_f$. Now we look at the case where ones wants to avoid a given spherical object *in space*, motionless in the considered referential, regardless of the speed. In other words, for a fixed choice of $(z_1, z_2) \in \mathbb{R}^2$, and $\varepsilon > 0$, we consider

$$\mathcal{Y}_f = \Big\{ (y_1, y_2, y_3, y_4) \in \mathbb{R}^4,\ \|(y_1 - z_1, y_2 - z_2)\|_{\mathbb{R}^2} \leq \varepsilon \Big\}. \tag{II.4.6}$$

In that case $\mathcal{Y}_f$ is unbounded; letting $z := (z_1, z_2, 0, 0)$, the support function of $\mathcal{Y}_f$ can be computed to be

$$\sigma_{\mathcal{Y}_f} : x \longmapsto \langle z, x \rangle + \varepsilon \|x\|_2 + \delta_{\{y \in \mathbb{R}^4,\ y_3 = y_4 = 0\}}(x),$$

where we use the convex analytic notation $\delta_C(x) = 0$ if $x \in C$ and $+\infty$ instead. We prove below a certified result for one such example.

Theorem II.17. Take $z_1 = z_2 = 0.5$, $\varepsilon = 0.1$, $M_2 = 1.15$, $M_\infty = 1$ and $T = 1$. Then $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from $(0, 0, 0, 0)^T$ in time T. Indeed, we find

$$J(p_f) \in [-0.1146, -0.0717], \quad \text{with} \quad p_f = (0.62, 0.78, 0, 0)^T.$$

Moreover, since $y_0 = (0, 0, 0, 0)^T$, for any $t \in [0, T]$, $\mathcal{Y}_f$ is not $\mathcal{U}$-reachable from $(0, 0, 0, 0)^T$ in time t.

II.4.4 Negative semi-definite Jacobi matrices

Control problem. In this section, we report on results for some randomly generated Jacobi matrices, with varying state dimensions n. That is, we consider matrices of the form

$$A = \begin{pmatrix} a_1 & c_1 & 0 & \dots & 0 \\ c_1 & a_2 & c_2 & \ddots & \vdots \\ 0 & c_2 & a_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \dots & 0 & c_{n-1} & a_n \end{pmatrix}. \tag{II.4.7}$$

with $a = (a_1, \dots, a_n) \in \mathbb{R}^n$, $c = (c_1, \dots, c_{n-1}) \in \mathbb{R}^{n-1}$.

These matrices are real symmetric, and to the best of our knowledge, no closed-form expressions are known for their eigenvalues and eigenfunctions, except in the specific case where the c_i's are all equal. Hence, for large values of n, diagonalising A becomes intractable. Even if it were accessible, it would be prone to numerical errors and we are not aware of any software that does produce such a diagonalisation within interval arithmetic.

We generate such a matrix in the following way: let $K > 0$ and $L > 0$. We draw the c_i's uniformly in $[-K, K]$. Then, we draw the a_i's uniformly in $(-2K - L, -2K]$. Thanks to the Gershgorin circle theorem, the resulting matrix is negative semi-definite.

We consider a single control u, thus $m = 1$. The corresponding matrix $B \in \mathbb{R}^{n\times 1}$ is $B = (1, \dots, 1)^T$. For a fixed $M > 0$, the constraint set is given by

$$\mathcal{U} = \{u \in \mathbb{R},\ |u| \leq M\},$$

for which we have $\sigma_{\mathcal{U}}(u) = M|u|$. The target y_f is chosen randomly, with i.i.d entries uniformly in $[-1, 1]$, then normalised such that $\|y_f\| = 0.05$.

Resolution method. Under the assumptions mentioned above, all eigenvalues of A are non-positive according to the Gershgorin circle theorem.

As a result, we are dealing with negative semi-definite matrices, enabling us to use estimates coming from Proposition II.14 upon using the Euler implicit scheme to approximate the matrix exponential.

Results. In the following example, we shall take $M = 1$, $T = 1$, $N_0 = 1{,}000$ and $y_0 = 0$.

Random targets. For each chosen dimension n, we generate 200 experiments with a target y_f of fixed norm $\|y_f\| = 0.05$, and a random matrix A drawn as explained previously (with $K = 2$, $L = 0.1$). More precisely, we run our descent algorithm to try and prove the non-$\mathcal{U}$-reachability of y_f from y_0 in time T. The following table shows the resulting means and standard deviations for the midpoint and size of the obtained intervals around $J(\frac{p_f}{\|p_f\|})$ where p_f is the last iterate of the optimisation algorithm.

Recalling the notation $E_{\mathcal{U}} = \{u \in E, t \in (0,T),\ u(t) \in \mathcal{U} \text{ for a.e. } t \in (0,T)\}$, it will be convenient to report values for the two terms involved in the definition of J, namely

$$J\Big(\frac{p_f}{\|p_f\|}\Big) = \sigma_{E_{\mathcal{U}}}\Big(L_T^* \frac{p_f}{\|p_f\|}\Big) - \Big\langle y_f, \frac{p_f}{\|p_f\|}\Big\rangle.$$

Indeed, once p_f and y_f are fixed, only the second term depends on the target y_f. Assume that $\langle y_f, \frac{p_f}{\|p_f\|}\rangle > 0$ and $J(\frac{p_f}{\|p_f\|}) > 0$ (so nothing is known about the $\mathcal{U}$-reachability of y_f). Then one can dilate y_f, i.e. change y_f to λy_f with $\lambda > 0$, and obtain a value $\lambda^\star > 0$ such that λy_f is not $\mathcal{U}$-reachable from 0 in time T, for any $\lambda > \lambda^\star$.

We also display below the proportion of targets which are proved to be non-reachable.

n	$\left\langle y_f, \frac{p_f}{\lVert p_f \rVert} \right\rangle$	Midpoints of $\sigma_{E_{\mathcal{U}}}\left(L_T^* \frac{p_f}{\lVert p_f \rVert}\right)$			Radii of $\sigma_{E_{\mathcal{U}}}\left(L_T^* \frac{p_f}{\lVert p_f \rVert}\right)$			Proportion of guaranteed non-reachable targets
		Mean	Median	Standard deviation	Mean	Median	Standard deviation	
2	0.0316	0.0027	0.0019	0.0063	0.0083	0.0082	0.0023	0.845
5	0.0369	0.0057	0.0036	0.0053	0.0148	0.0147	0.0026	0.91
10	0.0426	0.0043	0.0032	0.0033	0.0210	0.0209	0.0030	0.965
20	0.0457	0.0040	0.0035	0.0021	0.0288	0.0291	0.0032	0.97
50	0.0483	0.0067	0.0065	0.0026	0.0462	0.0461	0.0034	0.145

As the dimension n grows, one should expect that the proportion of final states y_f of norm $\lVert y_f \rVert = 0.05$ that are non-$\mathcal{U}$-reachable (from 0 in time T) should approach 1, as we keep a single control ($m = 1$). In fact, this is what is seen up until $n = 20$. Then, when the dimension is increased to $n = 50$, this proportion drops to about 15% if the tolerance δ defining the stopping criterion (II.4.1) is kept to its initial value $\delta = 10^{-5}$. By diminishing δ to $\delta = 5.10^{-7}$, we partially mitigate this problem, guaranteeing the non-$\mathcal{U}$-reachability of about 52% of states y_f. Obtaining even higher values becomes computationally prohibitive, making the case of dimension $n = 100$ intractable.

If, however, we increase the norms of targets y_f from 0.05 to 0.1, then a tolerance of $\delta = 10^{-5}$ is enough to certify that almost all such targets are non-$\mathcal{U}$-reachable, even in the case $n = 100$.

The main takeaway is that the tolerance δ should be adapted to the problem dimensions, and to how close the target of interest is to the unknown reachable set. Another degree of liberty is to increase the number of time steps N_0, which leads to a reduction of the error term at the expense of increased computation time. Regardless, these results suggest that our approach suffers from a sort of curse of dimensionality.

Size of the reachable set. We then try to estimate the size of the reachable set from $y_0 = 0$, i.e., $L_T E_{\mathcal{U}}$. For a given dimension n, we randomly choose a fixed matrix A in the same way as before. Then, we draw 1000 vectors $\tilde{p}_f$ at random on the unit sphere of $\mathbb{R}^n$, and report the statistics obtained for (the intervals) $\sigma_{L_T E_{\mathcal{U}}}(\tilde{p}_f) = \sigma_{E_{\mathcal{U}}}(L_T^* \tilde{p}_f)$.

n	Midpoints of $\sigma_{E_{\mathcal{U}}}(L_T^* \tilde{p}_f)$			Radii of $\sigma_{E_{\mathcal{U}}}(L_T^* \tilde{p}_f)$		
	Mean	Median	Standard deviation	Mean	Median	Standard deviation
2	0.3527	0.3874	0.1837	0.0087	0.0089	0.0014
5	0.3487	0.3112	0.2138	0.0138	0.0138	0.0022
10	0.3305	0.2783	0.2163	0.0204	0.0201	0.0027
20	0.2982	0.2523	0.2056	0.0292	0.0292	0.0032
50	0.3305	0.2839	0.2131	0.0467	0.0467	0.0037
100	0.3427	0.2983	0.2217	0.0651	0.0651	0.0032

As can be seen, although the midpoints of intervals are rather constant, the error term steadily increases, which leads to more difficult proofs of non-reachability. As already mentioned, one could increase N_0 to reduce errors.

II.5 Appendix: Minimal time for the streetcar example

Proposition II.18. Let $(x_f, y_f) \in \mathbb{R}^2$. The minimal time to steer system (II.4.2) from $(0,0)$ to (x_f, y_f) reads

$$T = \frac{-s y_f + 2\sqrt{\frac{1}{2} y_f^2 + s M x_f}}{M}, \quad \text{with } s = \operatorname{sign} f(x_f, y_f),$$

using the convention $\operatorname{sign}(0) = 0$, where $f : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(x, y) = x - \frac{1}{2M} y^2 \operatorname{sign}(y).$$

Proof. Let T be the optimal time steering system (II.4.2) from $(0,0)$ to (x_f, y_f). According to [LM86, Chapter 1], it is well-known that optimal controls are bang-bang equal a.e. to M or $-M$, with at most one switch, on the so-called switching locus defined by the implicit equation $f(x,y) = 0$.

More precisely, if $s < 0$, then the optimal control is $u = M\mathbb{1}_{(0,t_0)} - M\mathbb{1}_{(t_0,T)}$, where $t_0 \geq 0$ is the switching time, in other words the first time such that $f(x(t), y(t)) = 0$. Conversely, if $s > 0$, then $u = -M\mathbb{1}_{(0,t_0)} + M\mathbb{1}_{(t_0,T)}$. Easy but lengthy computations yield the following:

- If $f(x_f, y_f) = 0$, then for every $t \in [0,T]$, one has
$$y(t) = y_0 - Mt\,\mathrm{sign}(y_f) \quad \text{and} \quad x(t) = x_f - y_f t - \frac{1}{2}Mt^2\,\mathrm{sign}(y_f).$$

- Conversely, if $f(x_f, y_f) \neq 0$, then for every $t \in [0,T]$, one has
$$\begin{aligned} y(t) &= (-y_f - sMt)\mathbb{1}_{(0,t_0)} + (y_f + sM(t - 2t_0))\mathbb{1}_{(t_0,T)} \\ x(t) &= (x_f - y_f t - \tfrac{1}{2}sMt^2)\mathbb{1}_{(0,t_0)} + (x_f - y_f t + sM(\tfrac{1}{2}t^2 - 2t_0 t + t_0^2))\mathbb{1}_{(t_0,T)}. \end{aligned}$$

To conclude, it is important to notice that if $s \neq 0$, then $\mathrm{sign}(y(t_0)) = s$, which can be easily seen by distinguishing between several cases, depending on the sign of y_f and s.

To conclude, it remains to compute the switching time t_0. We claim that if $f(x_0, y_0) \neq 0$, then
$$t_0 = \frac{1}{M}\left(-s\,y_f + \sqrt{\frac{1}{2}y_f^2 + s\,Mx_f}\right).$$

Indeed, t_0 is characterised by the equation $f(x(t_0), y(t_0)) = 0$, which can be rewritten as the second order polynomial equation in the variable t_0:
$$0 = \left(x_f - s\,\frac{1}{2M}y_f^2\right) - y_f(1 + s^2)t_0 - s\,Mt_0^2.$$

Furthermore, the discriminant of this polynomial is positive. It follows that $y(T) = -y_f + sM(T - 2t_0)$, and therefore $T = \frac{s}{M}y_f + 2t_0$. The expected conclusion follows. □

III

Certified non-reachability for linear parabolic PDEs

Contents

Abstract

Analysing reachability associated to a control system is a subtle issue, especially for infinite-dimensional dynamics, and when controls are subject to bounded constraints. We develop a computer-assisted framework for establishing non-reachability in linear parabolic PDEs governed by strongly elliptic operators, extending recent finite-dimensional techniques introduced in [Has+24] to the PDE setting. The non-reachability of a given target is shown to be equivalent to proving that a properly defined dual functional takes negative values. Our approach combines rigorous numerics with explicit convergence estimates for discretisations of the adjoint equation, ensuring mathematically certified results with tight error bounds. We demonstrate the wide applicability of our framework on Laplacian-driven control systems, showcasing its accuracy and reliability under various types of control constraints.

III.1 Introduction

III.1.1 Non-reachability issues

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

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

Here X, the state space, and U, the control space, are two Hilbert spaces, $y_0 \in X$, $A : \mathcal{D}(A) \subset X \to X$ is an unbounded linear operator generating a C_0 semigroup over X, and $B : U \to X$ is a bounded linear operator.

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

More precisely, the methodology developed in this work applies to

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

$$\mathcal{U} \text{ is convex, closed and bounded.} \tag{III.1.1}$$

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

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

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

$$\begin{cases} \partial_t y - D\partial_{xx} y = \chi_\omega u, \\ y(t, 0) = y(t, 1) = 0. \\ y(0) = y_0. \end{cases} \tag{III.1.2}$$

with, say

$$D = 1, \quad y_0 = 0, \quad \omega = (\tfrac{1}{5}, \tfrac{2}{5}) \cup (\tfrac{4}{5}, 1), \quad \mathcal{U} = \{v \in L^2(\Omega),\ 0 \leq v \leq 1\,\text{a.e.}\}. \tag{III.1.3}$$

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

III.1.2 State of the art

Constraint-free controllability for linear PDEs has been the focus of extensive study from the 70's, leading to numerous significant results: general conditions, as the extension of Kalman's criterion for linear PDEs, or more specific results, such as the Geometric Control Condition for the wave equation [MZ04], the approximate controllability and exact null controllability of the heat equation [LR94] (see [MZ04; Boy22] for lecture notes compiling most common results). In recent years, some papers have studied constrained controllability, or reachability of linear ODEs and PDEs, uncovering general obstructions to controllability [DZ18; Bra72; LTZ21], due for example to comparison principles [LTZ18], or observability-based criteria caracterising the controllability of various problems with (un-)bounded and (a-)symmetric constraints on the control [Ber14; Ber19].

In this work, we focus on the reachability analysis of linear parabolic equations, typically with internal controls, a context that has been extensively studied, both in the unconstrained and constrained settings. For instance, for the reachable set of the heat equation with internal unconstrained control [KNT22; ELT22; HKT20; DE18; CR22]. Constraints on the state or the control have as well been considered, whether bounded or unbounded: [Ber20; Ant+24; CK22; Wan08; PTZ24].

Numerous methods exist in order to analyse reachability for finite-dimensional linear systems, including Hamilton-Jacobi type PDE formulations [MBT05; CT18], barrier functions [PJ04; KBH18], and set propagation (see the comprehensive review in [AFG21]). To the best of our knowledge, no such method has been designed in an infinite-dimensional setting, in such a way that it is possible to rigorously answer a question such as the one posed in Example III.2.

Our method is based on discretising an abstract necessary and sufficient condition with explicit bounds, and using interval arithmetic. Thus, our methodology belongs to the realm of computer-assisted proofs. It is worth noting that computer-assisted proofs based on interval arithmetic have already been applied to PDEs: see for example [Day+04; Bal+18; NPW19; Góm19; Kap+21; BBS24].

For a more comprehensive review of the literature, we refer the reader to Section 0.3.

III.1.3 Methodology and main results

III.1.3.a Separation argument

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

$$E_{\mathcal{U}} := \{u \in L^2(0,T;U), \text{ for a.e. } t \in (0,T),\, u(t) \in \mathcal{U}\}. \tag{III.1.4}$$

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

$$\forall\, u \in L^2(0,T;U), \quad L_T u = \int_0^T S_{T-t} B u(t)\, \mathrm{d}t. \tag{III.1.5}$$

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

$$\sup_{z \in L_T E_{\mathcal{U}}} \langle z, p_f \rangle < \langle y_f - S_T y_0, p_f \rangle \quad \Longleftrightarrow \quad \sup_{v \in E_{\mathcal{U}}} \langle v, L_T^* p_f \rangle_{L^2(0,T;U)} < \langle y_f - S_T y_0, p_f \rangle, \tag{III.1.6}$$

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

$$\forall\, p_f \in X, \quad J(p_f) := \sigma_{E_{\mathcal{U}}}(L_T^* p_f) - \langle y_f - S_T y_0, p_f \rangle = \int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f)\, \mathrm{d}t - \langle y_f - S_T y_0, p_f \rangle, \tag{III.1.7}$$

the above condition (III.1.6) amounts to $J(p_f) < 0$.

In fact, this sufficient condition becomes an equivalence under our assumption that $\mathcal{U}$ is a convex, closed and bounded case (and hence is convex and weakly compact). A separation argument outlined in detail in II.2, which goes through here when applied in the weak topology of X [Rud91, Theorem 3.4] (see as well Section I.2), leads to the following equivalence

$$y_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T > 0 \quad \Longleftrightarrow \quad \exists\, p_f \in X,\; J(p_f) < 0. \tag{III.1.8}$$

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

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

$$\mathcal{Y}_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T \quad \Longleftrightarrow \quad \exists\, p_f \in X,\ J(p_f;\mathcal{Y}_f) < 0. \qquad \text{(III.1.9)}$$

Remark III.4. We are assuming that $\mathcal{U}$ is closed convex and bounded, and that the control operator B is bounded. As can be checked, our approach works under the slightly more general assumptions that $B\mathcal{U} \subset X$ is closed, convex and bounded, and B is an admissible operator, see [TW09] for a definition of admissibility.

III.1.3.b Minimal times

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

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

$$\|y_f - S_T y_0\| > \sup_{t\in[0,T]} \|S_t\|_{\mathcal{L}(X)} T M_{B\mathcal{U}} \quad \Longrightarrow \quad y_f \text{ is not } \mathcal{U}\text{-reachable from } y_0 \text{ in time } T. \qquad \text{(III.1.10)}$$

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

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

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

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

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

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

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

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

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

III.1.3.c Control of discretisation and round-off errors

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

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

$$\begin{cases} p'(t) + A^* p(t) = 0, \\ p(T) = p_f. \end{cases} \tag{III.1.11}$$

With this notation in place, we may write

$$\begin{aligned} J(p_f) &= \int_0^T \sigma_{B\mathcal{U}}(S_{T-t}^* p_f)\,\mathrm{d}t - \langle y_f, p_f\rangle_X + \langle y_0, S_T^* p_f\rangle_X \\ &= \int_0^T \sigma_{B\mathcal{U}}(p(t))\,\mathrm{d}t - \langle y_f, p_f\rangle_X + \langle y_0, p(0)\rangle_X. \end{aligned} \tag{III.1.12}$$

In order to evaluate $J(p_f)$, not only does one need to compute a time integral and space integrals (the inner products in X), but more importantly, it is required to discretise the adjoint equation (III.1.11) both in time and in space – see Section III.2.2 for details. Hence, in practice we will only be able to evaluate a discretised function J_d, defined on a finite-dimensional subspace V_h.

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

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

For the first type of error, we derive in Theorem III.11 a fully explicit bound using functional analytic tools, while the second type is handled using interval arithmetic, thanks to the Matlab library INTLAB [Rum99].

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

$$J(p_f) \leq J_\mathrm{d}(p_{fh}) + e(p_{fh}, p_f) < 0, \tag{III.1.13}$$

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

Using this methodology, one can prove the non-reachability of targets for a large array of parabolic systems, which we develop in Section III.3. Let us now give an example of a certified result obtained with our method in the context of Example III.2, applied in the setting of parameters given by (III.1.3); in particular, recall that we have $y_0 = 0$, $T = 1$.

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

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

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

Extensions and perspectives Our approach relies on a few crucial hypotheses:

- convexity and compactness assumptions on the constraints and target set (III.1.1) and (III.3) to obtain the equivalence (III.1.8)

- continuity and coercivity of $-A^*$ (III.2.2) for small discretisation errors
- on a classical assumption of convergence of the family of discretisation spaces (see assumption $(\mathcal{V}_1)$).

A first range of potential extensions of our method would be to extend it under weaker assumptions. For instance, since finding p_f such that $J(p_f) < 0$ remains a sufficient condition for non-reachability, the method might still provide dual certificates for nonconvex target sets. Similarly, the boundedness assumptions of B and $\mathcal{U}$ could be combined into the weaker assumption of boundedness of $BE_{\mathcal{U}} = \{Bu, u \in E_{\mathcal{U}} \subset L^2(0,T,U)\}$. One could then consider unbounded constraints or an unbounded yet admissible operator $B \in L(U, \mathcal{D}(A^*)')$, at the cost of more complex formulae for $\sigma_{BE_{\mathcal{U}}}$ – either with a closed formula or with precise approximations.

One could also try weakening the continuity-coercivity hypotheses, which are key to upper-bound the discretisation errors with small constants regarding to the final time. This could involve restricting the search for dual certificates in a subspace where $-A^*$ is continuous and coercive, at the cost of equivalence (III.1.8), or ignore completely these hypotheses and try to manage constants growing exponentially with T, a major hindrance for computer-assisted proofs. This extension could for instance be done using different time-discretisation schemes (see Remark III.10), and could for instance allow changes of boundary conditions.

Other leads include major changes of the partial differential equation: an extension to non-homogeneous non-autonomous linear systems of the form $\partial_t y(t) = A(t)y(t) + B(t)u(t) + v(t)$ would require the use of resolvents to obtain (less accurate) discretisation error bounds. A significantly bolder perspective would be tackling semilinear or nonlinear parabolic equations. Our approach being fundamentally reliant on the linearity of L_T to compute $\sigma_{L_T E_{\mathcal{U}}}$, this would require significant modifications, perhaps necessitating proving inclusions of the reachable set in those of an appropriately chosen linear PDE.

Outline of the article. In Sections III.2.1 and III.2.2, we provide details about the functional analytic framework of the method, as well as the different hypotheses we make both on the operator A and on the discretisation methods. Proposition III.9 outlines the error bounds associated to discretising the adjoint equation (III.1.11). Section III.2.3 contains the main theoretical result, that is, Theorem III.11 which provides precise discretisation errors incurred when approximating the dual functional J mentioned in (III.1.13) by its discretised counterpart J_{d}. Finally, Section III.2.4 summarises the whole method and hypotheses.

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

III.2 Key results and proofs

III.2.1 Functional analytic framework

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

Let X and $V \subset X$ be two complex Hilbert spaces, with V densely and continuously embedded into X, equipped with the norms $\|\cdot\|_X$ (associated to the inner product $\langle \cdot, \cdot \rangle_X$) and $\|\cdot\|_V$. For ease of readability, we will drop the subscript X when dealing with $\|\cdot\|_X$ and $\langle \cdot, \cdot \rangle_X$. Let us identify X and its dual X', with the associated Gelfand triple $V \subset X \subset V'$.

Remark as well that even though we need complex Hilbert-space to make use of operator and semigroup theory, we have defined the separation argument (III.1.8) over a real Hilbert-space. This separation argument would stand in a complex Hilbert-space – at the cost of considering real parts of every scalar product. Similarly to the convention used in Chapter I (see Page 32), as all our examples in Section III.3.3 are real-valued, there we shall consider real-Hilbert spaces.

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

$$\mathcal{D}(\mathcal{A}) = \{x \in V, \mathcal{A}x \in X\}. \tag{III.2.1}$$

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

$$\forall\, (v, w) \in \mathcal{D}(\mathcal{A}) \times V, \quad \begin{cases} |\langle \mathcal{A}v, w\rangle| \leq a_1 \|v\|_V \|w\|_V \\ \operatorname{Re}\left(\langle \mathcal{A}v, v\rangle\right) \geq a_0 \|v\|_V^2. \end{cases} \tag{III.2.2}$$

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

$$\forall\, v \in \mathcal{D}(\mathcal{A}), \quad \langle \mathcal{A}v, v\rangle \in \mathcal{S}_\alpha := \{z \in \mathbb{C},\ z = 0 \text{ or } |\arg z| \leq \alpha\}, \tag{III.2.3}$$

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

$$\forall\, z \notin \mathcal{S}_\alpha, \quad z\operatorname{Id} - \mathcal{A} : \mathcal{D}(\mathcal{A}) \to X \text{ is an isomorphism.} \tag{III.2.4}$$

The combination of (III.2.3) and (III.2.4) corresponds to $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ being a so-called *mα-accretive operator*. We shall need to use functions of such operators: if r is a rational fraction, bounded on $\mathcal{S}_\alpha$, written in the form

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

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

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

This definition then straightforwardly extends to functions f that may be written as the uniform limit of such rational fractions in $\mathcal{S}_\alpha$ – see Section I.1 for details about functions of operators. For this class of functions, we have the important following estimate.

Theorem III.8 ([CD03])**.** For any function $f : \mathcal{S}_\alpha \to \mathbb{C}$ that is the uniform limit of bounded rational fractions on $\mathcal{S}_\alpha$,

$$\|f(\mathcal{A})\|_{\mathcal{L}(X)} \leq C_\alpha \sup_{z\in\mathcal{S}_\alpha} |f(z)|, \tag{III.2.5}$$

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

It is commonly known that the opposite of a mα-accretive operator generates a C_0 semigroup (see Remark I.11). Furthermore, the second inequality of (III.2.2) implies that $\mathcal{A}$ is an isomorphism from V to V' by the Lax-Milgram theorem, with inverse $\mathcal{A}^{-1} : V' \to V$. Finally, we may define the adjoint $\mathcal{A}^* \in \mathcal{L}(V', V)$ of $\mathcal{A}$.

III.2.2 Discretisation errors

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

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

$$\forall\, v_h, w_h \in V_h, \quad \mathcal{A}_h v_h \in V_h \quad \text{and} \quad \langle \mathcal{A}_h v_h, w_h\rangle = \langle \mathcal{A}v_h, w_h\rangle. \tag{III.2.6}$$

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

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

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

$$\begin{cases} z'(t) = -\mathcal{A}z(t) \\ z(0) = z_0. \end{cases} \tag{III.2.7}$$

Considering Euler's implicit scheme for the corresponding discrete problem (through V_h), we let $N_0 \in \mathbb{N}^*$, $\Delta t := \frac{T}{N_0}$ and for $z_{h,0} \in V_h$ consider

$$\forall\, n \in \{0,\dots,N_0\}, \quad z_{h,n} = (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n}\, z_{h,0}. \tag{III.2.8}$$

The associated error can be estimated is as follows.

Proposition III.9. Assume that the couple given by $(\mathcal{A},\mathcal{D}(\mathcal{A}))$ and V_h satisfies (III.2.2)-($\mathcal{V}_1$), and let $z_0 \in \mathcal{D}(\mathcal{A})$, $z_{h,0} \in V_h$. Then, letting $z : [0,T] \to X$ be the solution to (III.2.7) and $z_{h,n}$, $n \in \{0,\dots,N_0\}$ be defined by (III.2.8), we have

$$\forall\, n \in \{0,\dots,N_0\}, \quad \|z(t_n) - z_{h,n}\| \leq C_\alpha \|z_0 - z_{h,0}\| + \Big(C_2 h^2 + C_3 \Delta t\Big) \|\mathcal{A}z_0\|. \tag{III.2.9}$$

where

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

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

Remark III.10. In this article, we have only considered Euler's implicit time-discretisation scheme to approximate (III.2.7). A more general class of schemes consists in using Euler's implicit scheme to discretise

$$\begin{cases} w'(t) = \kappa w(t) - \mathcal{A}w(t) \\ w(0) = z_0, \end{cases} \tag{III.2.10}$$

and then notice that $z(t) = e^{-\kappa t} w(t)$. This could have two main advantages: firstly, if $\mathcal{A}$ is a_0-coercive and a_1-continuous, and if V is continuously embedded in X with constant c, then $\mathcal{A} - \kappa\,\mathrm{Id}$ is $(a_1 + c^2\kappa)$-continuous, and if $0 < \kappa < \frac{a_0}{c^2}$ then $\mathcal{A} - \kappa\,\mathrm{Id}$ is $(a_1 - c^2\kappa)$-coercive. Therefore one can apply Proposition III.9 to (III.2.10) and obtain discretisation errors decreasing exponentially with time.
On another hand, these schemes could help extending the method to non-coercive operators: if $\mathcal{A} - \kappa\,\mathrm{Id}$ is coercive for $\kappa < 0$, then the same trick allows to use Proposition III.9 and obtain explicit discretisation errors – admittedly, those increase exponentially in time and would therefore be of little use even on small timescales.
Notice as well that we could consider spectral methods: in the rare case where one has access to eigenvalues and eigenvectors of $\mathcal{A}$, then one could avoid all discretisation errors related to the discretisation of (III.2.7).

III.2.3 Control problem

We are now given the control problem ($\mathcal{S}$), a family of finite-dimensional subspaces V_h (of dimension denoted M_h) indexed by $h > 0$. The important assumptions for us will concern the unbounded operator $\mathcal{A} := -A^*$, with domain $\mathcal{D}(A^*)$ and the finite dimensional subspaces V_h:

- $(\mathcal{A},\mathcal{D}(\mathcal{A}))$ satisfies (III.2.2),

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

As before, we introduce the corresponding notation $\mathcal{A}_h$, defined by the relation (III.2.6).

We recall the assumption (III.1.1) that $\mathcal{U}$ is closed, convex and bounded. We let $M_{B\mathcal{U}} > 0$ be defined by

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

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

Starting from the equation (III.1.12) (and with the change of variable $t = T - t$ within the integral), we have

$$\forall p_f \in X, \quad J(p_f) = \int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f)\,\mathrm{d}t - \langle y_f, p_f\rangle_X + \langle y_0, S_T^* p_f\rangle_X. \tag{III.2.11}$$

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

$$J_{\Delta t,h}(p_{fh}) = \Delta t \sum_{n=1}^{N_0} \sigma_{B\mathcal{U}}((\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh}) - \langle y_f, p_{fh}\rangle + \left\langle y_0, (\mathrm{Id} + \Delta t\mathcal{A}_h)^{-N_0} p_{fh}\right\rangle \tag{III.2.12}$$

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

Theorem III.11. Assume that $(-A^*, \mathcal{D}(A^*))$ and the family V_h satisfy (III.2.2) as well as $(\mathcal{V}_1)$. Then for all $p_f \in \mathcal{D}(A^*)$, $p_{fh} \in V_h$, we have

$$\begin{aligned} |J(p_f) - J_{\Delta t,h}(p_{fh})| \leq \Big(\tfrac{1}{2} M_{B\mathcal{U}} T \Delta t + (\|y_0\| + M_{B\mathcal{U}} T)(C_2 h^2 + C_3 \Delta t)\Big) \|A^* p_f\| \\ + \left((\|y_0\| + M_{B\mathcal{U}} T)\, C_\alpha + \|y_f\|\right) \|p_f - p_{fh}\|. \end{aligned} \tag{III.2.13}$$

Proof. We decompose the error in the form

$$|J(p_f) - J_{\Delta t,h}(p_{fh})| \leq \underbrace{|J(p_f) - \tilde{J}(p_f)|}_{(I)} + \underbrace{|\tilde{J}(p_f) - J_{\Delta t,h}(p_{fh})|}_{(II)}, \tag{III.2.14}$$

where

$$\forall p_f \in X, \quad \tilde{J}(p_f) := \Delta t \sum_{n=1}^{N_0} \sigma_{B\mathcal{U}}(S_{t_n}^* p_f) - \langle y_f, p_f\rangle + \langle y_0, S_T^* p_f\rangle. \tag{III.2.15}$$

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

$$\left| \int_0^T f(t)\,\mathrm{d}t - \Delta t \sum_{k=1}^{N_0} f(k\Delta t) \right| \leq \frac{1}{2} K T \Delta t. \tag{III.2.16}$$

Our aim is now to apply the above estimate to $f : t \mapsto \sigma_{B\mathcal{U}}(S_t^* p_f)$. Since $B\mathcal{U}$ is bounded (by $M_{B\mathcal{U}} > 0$), $\sigma_{B\mathcal{U}}$ is $M_{B\mathcal{U}}$-Lipschitz continuous (recall Proposition I.18), hence we have

$$|f(t) - f(s)| \leq M_{B\mathcal{U}} \|S_t^* p_f - S_s^* p_f\|. \tag{III.2.17}$$

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

$$|f(t) - f(s)| \leq M_{B\mathcal{U}} \sup_{t\in[0,T]} \|A^* S_t^* p_f\|\; |t-s| \leq M_{B\mathcal{U}} \|A^* p_f\|\; |t-s|, \tag{III.2.18}$$

where we used $A^* S_t^* p_f = S_t^* A^* p_f$ and, given that $-A^*$ is mα-accretive, the bound $\|S_t^*\|_{\mathcal{L}(X)} \leq 1$ for all $t \geq 0$. Summing up, the above Lipschitz estimate entails

$$(I) = |J(p_f) - \tilde{J}(p_f)| \leq \frac{1}{2} \Delta t M_{B\mathcal{U}} T \|A^* p_f\| \tag{III.2.19}$$

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

$$|\tilde{J}(p_f) - J_{\Delta t,h}(p_{fh})| \leq \Delta t \sum_{n=1}^{N_0} |\sigma_{B\mathcal{U}}(S^*_{t_n} p_f) - \sigma_{B\mathcal{U}}((\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh})| \tag{III.2.20}$$

$$+ \left| \left\langle y_0, S^*_T p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-N_0} p_{fh} \right\rangle \right| + \|y_f\| \|p_f - p_{fh}\|. \tag{III.2.21}$$

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

$$\begin{aligned} \left| \left\langle y_0, S^*_T p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-N_0} p_{fh} \right\rangle \right| &\leq \|y_0\| \|S^*_T p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-N_0} p_{fh}\| \\ &\leq \|y_0\| \left(C_\alpha \|p_f - p_{fh}\| + \left(C_2 h^2 + C_3 \Delta t \right) \|A^* p_f\| \right). \end{aligned}$$

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

$$\begin{aligned} \Delta t \sum_{n=1}^{N_0} |\sigma_{B\mathcal{U}}(S^*_{t_n} p_f) - \sigma_{B\mathcal{U}}((\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh})| &\leq \Delta t \, M_{B\mathcal{U}} \sum_{n=1}^{N_0} \|S^*_{t_n} p_f - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n} p_{fh}\| \\ &\leq M_{B\mathcal{U}} T \left(C_\alpha \|p_f - p_{fh}\| + \left(C_2 h^2 + C_3 \Delta t \right) \|A^* p_f\| \right). \end{aligned}$$

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

□

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

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

Remark III.13. Notice that the time-discretisation constants may be suboptimal in both Proposition III.9 and Theorem III.11. Indeed, they have been computed under the only assumption that $\mathcal{A}$ is a mα-accretive operator, and not accounting for its coercivity. This might induce an upper-bound of the form: for all $t \geq 0$, $\|S^*_t\| \leq \gamma e^{-\varepsilon t}$, with $\gamma \geq 1, \varepsilon > 0$, which would tighten the estimates of (III.2.17- III.2.18) (see Remark III.10 for a lead on obtaining those estimates). As for the discretisation errors on the adjoint equation (III.2.7), the estimates can be traced back through Proposition III.9 to (III.4.15), where the supremum of the implicit Euler scheme is taken on $\mathcal{S}_\alpha$. Instead, it could be taken on the numerical range, defined as

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

However, although the literature about this set is extensive (see for instance [CD03; CP17; CG19] and the references therein), forfeiting the sectorial simplification might increase the value of the equivalent of the constant C_α in Theorem III.8.

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

- The first one is to rely on simple and thoroughly studied discretisation subspaces, such as the finite element method. Numerical computations and estimates of constants are made easier, but we typically do not have $p_{fh} \in \mathcal{D}(A^*)$. Hence, we must interpolate p_{fh} in some way to get $p_f \in \mathcal{D}(A^*)$, which in turn requires to appropriately bound $\|p_f - p_{fh}\|$.
- The second one is to choose $V_h \subset \mathcal{D}(A^*)$ (for instance, splines) so that we may simply set $p_f = p_{fh}$, thus circumventing the interpolation problem altogether. The

price to pay is that the computation of mass and stiffness matrices is much more complex. Also, a larger dimension for the subspaces V_h is required to get the same discretisation parameter h, which results in increased computational costs.

The examples studied in Section III.3.3 will be made using the first approach, with $\mathbb{P}_1$ finite elements. We have also considered the second method using cubic splines (see Subsubection I.3.2.b for details about spline discretisation spaces): this is a more challenging method, theoretically first and especially numerically, which has not yet yielded better results than the first one. To improve it, one might assume more regularity (for instance, $\mathcal{D}((A^*)^2)$) of the initial condition in Proposition III.9, to obtain a higher order space convergence. To tackle this, we have proved in Subsubection I.3.2.c approximation results, which in turn require interpolation of cubic splines into more regular functions – for instance, the heptic splines introduced in Subsubsection I.3.2.d.
Howver, such a new interpolation method would face two main hindrances: first, the time-discretisation convergence is limited to order 1 (recall that we also have to approximate the integral of the Lipschitz functional $\sigma_{\mathcal{U}}$), which somewhat mitigates the impact of a higher order of convergence with respect to space. On another note, this method requires much heavier numerical computations. This, as we will see in Section III.3, will lead to high rounding errors, which quickly overwhelm discretisation errors.

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

$$I_h : z \in \mathbb{R}^{M_h} \mapsto \sum_{i=1}^{M_h} z_i\, \psi_i \in V_h. \tag{III.2.22}$$

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

$$\mathcal{M}_{i,j} = \langle \psi_i, \psi_j \rangle \quad \text{and} \quad \mathcal{K}_{i,j} = \langle \mathcal{A}\psi_j, \psi_i \rangle. \tag{III.2.23}$$

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

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

$$J_{\Delta t,h}(I_h z) = \Delta t \sum_{n=1}^{N_0} \sigma_{B\mathcal{U}}(I_h(\mathrm{Id} + \Delta t \mathcal{M}^{-1}\mathcal{K})^{-n} z) - \langle y_f, I_h z \rangle + \left\langle y_0, I_h(\mathrm{Id} + \Delta t \mathcal{M}^{-1}\mathcal{K})^{-N_0} z \right\rangle . \tag{III.2.24}$$

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

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

III.2.4 Methodology

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

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

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

We first need to check that $(-A^*, \mathcal{D}(A^*))$ (with $\mathcal{D}(A^*) \subset V$, V densely and continuously embedded in X) satisfies (III.2.2), where we must explicitly compute a_0, a_1.

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

For a suitable basis $(\psi_1, \dots, \psi_{M_h})$ of V_h, we compute the mass and stiffness matrices associated to the discretisation method, given by (III.2.23).

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

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

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

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

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

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

In both cases, we end up with a bound

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

where C is known explicitly.

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

III.3 Application to 1D parabolic problems

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

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

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

III.3.1 Preliminaries on discretisation methods

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

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

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

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

Let us gather the few main results needed regarding $\mathbb{P}_1$ finite elements – we refer for instance to [QSS06, Section 3] for a more precise setup. First, in view of the fact that the main operator of interest will be the Dirichlet Laplacian, we shall recall the following standard estimates with explicit constants. For the proof, we refer to Proposition I.51.

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

$$\left\| f - \sum_{i=1}^{N_h - 1} f(x_i)\, \psi_i \right\|_{H_0^1} \leq \frac{h}{\sqrt{2}} \|f''\|_{L^2}, \tag{III.3.1}$$

and

$$\left\| f - \sum_{i=1}^{N_h - 1} f(x_i) \psi_i \right\|_{L^2} \leq \frac{h^2}{2\sqrt{2}} \|f''\|_{L^2}. \tag{III.3.2}$$

Interpolation using cubic splines

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

As already explained, it will be necessary to interpolate this function to build $p_f \in \mathcal{D}(\Delta)$ so as to make use of Theorem III.11. Such an interpolation should be done while making the discretisation error negligible. For a given choice of discretisation parameters Δt, and h, this will be all the more likely that both terms $\|p_f - p_{fh}\|$ and $\|A^* p_f\|$ are small. If $A = \Delta$, then by the estimate (III.3.2), $\|p_f - p_{fh}\|_{L^2}$ can be controlled by $\|p_f''\|_{L^2}$.

All in all, a natural requirement is to interpolate p_{fh} into p_f in a way that makes $\|p_f''\|_{L^2}$ as small as possible. Using the following Lemma (see [De 01, Chapter V, Theorem (5-7)] for details), we are then led to using cubing splines – see also Section I.3 for details about spline interpolation.

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

$$\inf_{\substack{f \in H^2(0,1) \\ \forall\, i \in \{0, \dots, N_h\},\, f(x_i) = q_i}} \|f''\|_{L^2}. \tag{III.3.3}$$

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

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

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

Given the discretised functional $J_{\Delta_t,h}$ defined by (III.2.12), a key step is to efficiently find $p_{fh} \in V_h$ at which it takes negative values, if it ever exists. This can be a computationally challenging step that must consequently be carried out with care. First, it is critical that this step can be done *completely independently* from certifying the negativity of $J(p_f)$: any method and acceleration available to minimise $J_{\Delta t,h}$ can (and should) therefore be used.

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

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

Fenchel duality context

First, let us notice that one can address the question of finding $p_f \in X$ by considering the optimisation problem

$$d := \inf_{p_f \in X} J(p_f) = \inf_{p_f \in X} \sigma_{E_{\mathcal{U}}}(L_T^* p_f) - \langle y_f, p_f \rangle + \langle y_0, S_T^* p_f \rangle. \tag{III.3.4}$$

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

Now, as pointed out in Section I.2, this optimisation problem is the Fenchel-dual to the following primal problem

$$\pi := \inf_{u \in L^2(0,T;U)} \delta_{E_{\mathcal{U}}}(u) + \delta_{\{y_f - S_T y_0\}}(L_T u), \tag{III.3.5}$$

where δ_C is the convex indicator of a set C (taking the value 0 in C, and $+\infty$ outside of C). By our previous arguments, we have $d = -\pi$; such a strong duality can also be proved by using the Fenchel-Rockafellar theorem.

This duality structure allows for efficient minimisation algorithms, such as the Chambolle-Pock primal-dual algorithm [CP11]. For appropriately chosen step sizes, they are known to converge to saddle points of the associated saddle point functional whenever such points exist. Remark that in practice convergence of such algorithms would depend on the strong duality of the discretised primal and dual problems: this is linked to the reachability analysis of the discretised control problem – which is a complex question in itself, see for instance [Boy13].

However, notice that in the case of non-reachability, the saddle-point of the continuous primal-dual problem does not exist; thus the algorithm will not converge, and the norm of the iterates might diverge. For reasons we will evoke later in this section, this is not an issue (see (III.3.7)), and we shall therefore not delve into a convergence analysis of the algorithm in the discrete setting.

Consequently, some stopping criterion must be chosen in order to converge onto a p_{fh}. In practice, two such criteria are used in the numerical experiments: a numerical check of whether $J_{\Delta t,h}$ seems negative enough, and another of whether the algorithm is still moving, in the sense that $\|\frac{p_{i+1}}{\|p_{i+1}\|} - \frac{p_i}{\|p_i\|}\|$ is small enough. The result of that 'minimisation' step will abusively be called 'minimiser of $J_{\Delta t,h}$'.

Regularisation

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

$$\forall\, p_f \in \mathcal{D}(A^*), \quad J_\lambda(p_f) = J(p_f) + \frac{\lambda}{2}\|A^* p_f\|^2, \tag{III.3.6}$$

where $\lambda > 0$ can be thought of as a regularisation parameter. In that case, since A^* satisfies (III.2.2) and V is continuously embedded in X, J_λ is strongly convex and thus we have the existence and uniqueness of a minimiser of J_λ on $\mathcal{D}(A^*)$. Furthermore, using the Chambolle-Pock algorithm, one attains a convergence rate towards the saddle-point of $\mathcal{O}(\frac{1}{k})$. The major advantage of choosing such a regularisation term can be seen in Theorem III.11. Indeed, the computed-assisted proof of non-reachability essentially boils down to

$$\exists\, (p_{fh}, p_f) \in V_h \times \mathcal{D}(A^*), \quad \frac{J_{\Delta t,h}(p_{fh})}{\|A^* p_f\|} < -C, \tag{III.3.7}$$

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

As it turns out, the following lemma proves that the choice of λ does not influence the direction of the minimiser. We refer the reader to Lemma I.28 for its proof.

Lemma III.17. Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be two convex continuous functions such that

$$\forall\, \alpha \geq 0, \forall\, x \in X \quad \begin{cases} f(\alpha x) = \alpha f(x) \\ g(\alpha x) = \alpha^2 g(x). \end{cases} \tag{III.3.8}$$

Assume furthermore that g is positive outside of 0. Then,

$$\inf_{x\in X} f(x) + \frac{\lambda}{2} g(x) = -\frac{1}{2\lambda} \sup_{s\in S_g} \left(\min(f(s), 0)\right)^2, \tag{III.3.9}$$

and if the sup is reached, then:

$$\exists\, p \in X, \forall\, \lambda > 0, \exists\, r > 0, \quad f(rp) + \frac{\lambda}{2} g(rp) = \inf_{x\in X} f(x) + \frac{\lambda}{2} g(x). \tag{III.3.10}$$

Therefore the choice of $\lambda > 0$ has no influence over the direction of J_λ's 'minimiser'. Note however that choosing a 'reasonable value' of λ may improve the theoretical or numerical convergence rate of the minimisation algorithm being used. Indeed, for $\lambda > 0$, J_λ is λa_0 strongly convex, one can use an accelerated version of Chambolle-Pock's algorithm (see [CP11] for details) and obtain better convergence rates. In that case, the convergence of the iterations $(p^k)_{k\in\mathbb{N}}$ of the algorithm towards a minimiser $p^\star$ essentially satisfies the following statement: for a constant $C > 0$ depending on parameters of the problem and for $k \in \mathbb{N}$,

$$\|p^k - p^\star\| \leq \frac{C}{\lambda a_0 k}. \tag{III.3.11}$$

III.3.3 Examples of computer-assisted proofs

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

III.3.3.a 1D heat equation

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

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

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

First, it is classical that $-A^*$ is selfadjoint, namely $-A^* = -A$ with domain $H^2(0,1) \cap H_0^1(0,1)$, and that it satisfies (III.2.2) with $a_0 = a_1 = 1$, and hence is mα accretive with $\alpha = 0$.

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

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

$$y_0 = 0, \quad \omega = \Omega, \quad \mathcal{U} = \{u \in U,\ \|u\|_U \leq 1\}. \tag{III.3.12}$$

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

$$\forall\, v \in X, \quad \sigma_{B\mathcal{U}}(v) = \|v\|_X. \tag{III.3.13}$$

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

Lemma III.18. The following statements are equivalent:

- $\lambda\psi_k$ is $\mathcal{U}$-reachable from 0 in time $T > 0$
- There holds
$$|\lambda| \leq \lambda_k^\star := \sqrt{2} M_{B\mathcal{U}} \frac{1 - e^{-k^2\pi^2 T}}{k^2\pi^2}.$$

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

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

Moreover, from the constraints (III.3.12), we have that for all $t \in [0,T]$, $\|u(t)\| \leq 1$. Thus, for all $t \in [0,T]$, $|\alpha(t)| \leq \sqrt{2} M_{B\mathcal{U}}$, which implies that

$$|\lambda| \leq \sqrt{2} M_{B\mathcal{U}} \int_0^T e^{-k^2\pi^2(T-t)} \,\mathrm{d}t = \sqrt{2} M_{B\mathcal{U}} \frac{1 - e^{-k^2\pi^2 T}}{k^2\pi^2} =: \lambda_k^\star. \tag{III.3.14}$$

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

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

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

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

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

As we can see, the method is very effective for small frequencies, but its precision decreases quickly on higher frequencies. This can be easily understood since the error term in Theorem III.11 depends on the second derivative of the interpolant, and thus increases with the square of the frequency. Here, no minimisation process was required to find an optimal dual certificate: indeed, one can show that in this context, if $y_f = \lambda\psi_k$, $p_f = \frac{y_f}{\|y_f\|}$ satisfies $J(p_f; y_f) < 0$ if and only if y_f is not reachable. Actually, one can show that

$$\forall\, q \in \mathcal{D}(A), \quad \langle q, \psi_k\rangle = 0 \implies J(p_f + q; y_f) \geq J(p_f; y_f), \tag{III.3.15}$$

meaning that $p_f = y_f = \psi_k$ (or any positive multiple thereof) is an optimal dual certificate of non-reachability. Let us emphasise that this is only the case because of the simple control constraints, and especially because $B = \mathrm{Id}$. This result is confirmed by numerical experiments: when minimising $J_{\Delta t,h}(\cdot; \psi_k)$, the algorithm stabilises around the direction given by ψ_k very quickly.

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

$$y_0 = 0, \quad \omega = (\tfrac{1}{5}, \tfrac{2}{5}) \cup (\tfrac{4}{5}, 1), \tag{III.3.16}$$

and

$$\mathcal{U} = \{u \in U,\ 0 \leq u(x) \leq 1 \ \text{ for a.e. } \ x \in [0,1]\}. \tag{III.3.17}$$

In this setting, note that $\|B\| = M_{\mathcal{U}} = 1$, but one can easily compute that $M_{B\mathcal{U}} = \sqrt{|\omega|} = \sqrt{\frac{2}{5}}$, so that $M_{B\mathcal{U}} < M_{\mathcal{U}}\|B\| = 1$. Letting $z_+ = \max(z, 0)$ for $z \in \mathbb{R}$, we compute

$$\forall\, v \in X, \qquad \sigma_{B\mathcal{U}}(v) = \sup_{u\in\mathcal{U}} \langle u, \chi_\omega v\rangle = \int_\omega \sup_{0\leq y\leq 1} y\, u(x)\,\mathrm{d}x = \int_\omega u_+(x)\,\mathrm{d}x. \tag{III.3.18}$$

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

- First, the crude inclusion of the reachable set in a ball of centre $S_T y_0$ based on Lemma III.5 shows that $\|y_f - S_T y_0\| \leq M_{B\mathcal{U}} T$ is a necessary condition for y_f to be $\mathcal{U}$-reachable from 0 in time T.
- Second, the parabolic comparison principle leads to $S_T y_0 \leq y(T) = S_T y_0 + L_T u \leq S_T y_0 + L_T \overline{u}$ for all controls taking values in $\mathcal{U}$, where $\overline{u}(t,x) = 1$ for a.e. $t \in (0,T)$, $x \in (0,1)$. In the case where $y_0 = 0$, this yields a necessary condition for a target y_f to be $\mathcal{U}$-reachable from 0 in time $T > 0$, given by
$$0 \leq y_f \leq L_T \overline{u}. \tag{III.3.19}$$

One can easily prove that for this ω (or for any union of two disjoint intervals), one can approximate $L_T\bar{u}$ to obtain the following error estimate

$$\forall\, n \in \mathbb{N}^*, \qquad \left\| L_T\bar{u} - \sum_{k=1}^{n} \left[\frac{1-e^{-Tk^2\pi^2}}{k^2\pi^2} \int_\omega \frac{\psi_k}{\|\psi_k\|} \right] \frac{\psi_k}{\|\psi_k\|} \right\|_{L^\infty([0,1])} \leq \frac{4}{\pi^3 n^2}, \tag{III.3.20}$$

which allows for proofs of non-reachability using the parabolic comparison principle. Therefore, to emphasise the usefulness of our method, all the computer-assisted results of non-reachability we will provide below satisfy (III.3.19), at least numerically.

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

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

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

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

Corollary III.20. For the system (S) with operator B associated to ω given by (III.3.16) and constraints (III.3.17), the target $\lambda\psi_1$ is not $\mathcal{U}$-reachable for $T = 1$ if $\lambda \geq 0.035$. Indeed, using the dual certificate p_f shown in Fig III.2, one can show that

$$J(p_f; 0.035\,\psi_1) \in [-0.000718, -0.000111] < 0. \tag{III.3.21}$$

Notice that, unlike the result proved in (III.3.15), the dual certificate p_f differs widely from the target y_f (see Fig III.2). This is mainly due to the presence of the operator $B = \chi_\omega$ that introduces a higher space heterogeneity. The optimisation step tends to create a dual certificate negative on ω, so as to reduce the value of $\sigma_{E_{\mathcal{U}}}(L_T^* p_f)$, while maximising $\langle y_f, p_f \rangle$, which translates in maximising the values of p_f on $[0,1]\backslash\omega$, since here $y_f = \psi_1 \geq 0$.

The result proved in Corollary III.20 happens to be near the limit of what can be done with the computer-assisted proofs developed in this thesis, in the sense that the very fine discretisation allowed for very small discretisation errors (approximately $2.72 \cdot 10^{-4}$). However, a more precise discretisation comes at the cost of higher computational costs (here, several hours on a standard desk computer), and often with higher rounding errors: here, they amounted to $3.13 \cdot 10^{-5}$. We can see that there is a turnpoint between discretisation and rounding errors leading to an increase of the total error.

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

Corollary III.21 (of Lemma III.6)**.** Let $y_f = \frac{1}{50}\psi_1$. With constraints (III.3.16) and (III.3.17), the minimal time $T^\star \in (0, +\infty]$ needed to steer (S) from $y_0 = 0$ to y_f satisfies $T^\star \geq T = 0.13$. Indeed, we have that

$$J(p_f; T) \in [-0.000192, -0.0000268] < 0, \tag{III.3.22}$$

where p_f is plotted in Fig III.2.

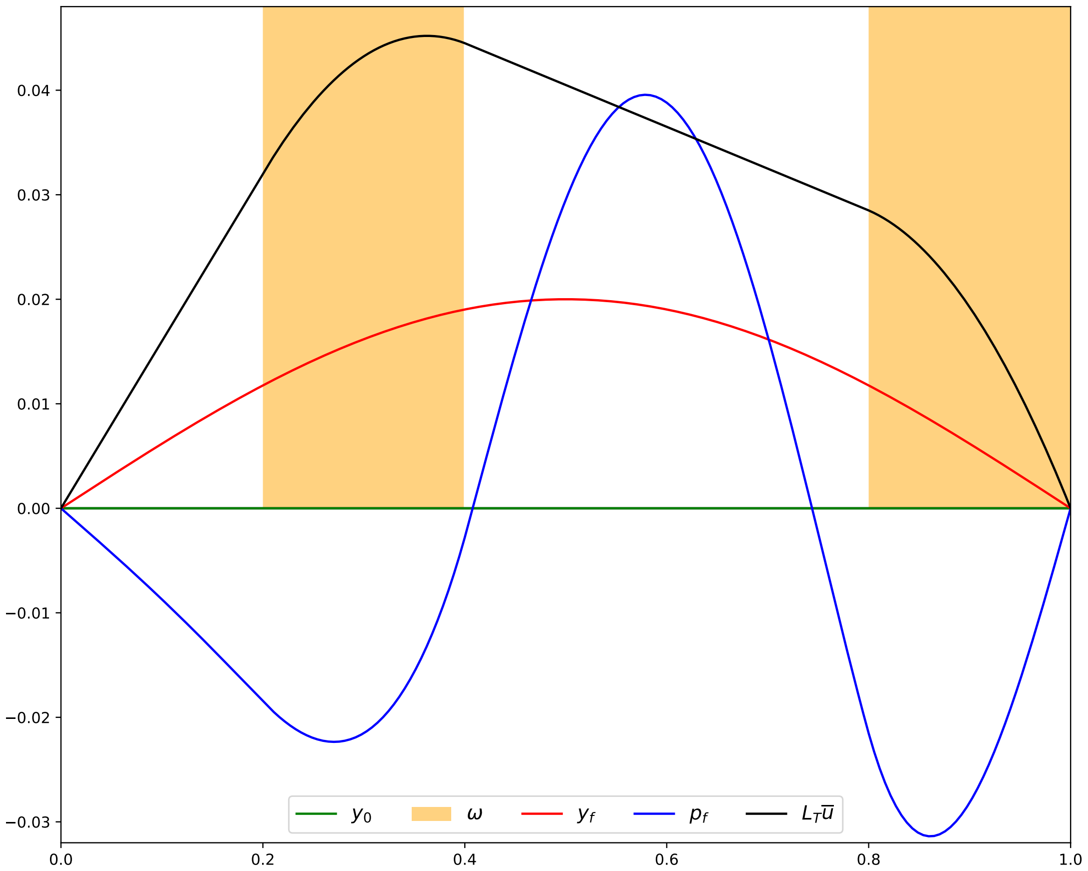


Figure III.2: Target and optimal dual certificate for Proposition III.21.

The method can also be applied to less smooth targets, such as absolute value based functions (see Fig III.3). However, minimising $J(\cdot; y_f)$ might provide less smooth dual certificates p_{fh}, which would lead to huge $\|A^* p_f\|$ after interpolation into p_f. Indeed, as stated before, the minimiser p_f will try to maximise the scalar product $\langle y_f, p_f \rangle$, and therefore mimic the regularity of y_f. To circumvent this, one might minimise the regularised functional introduced in (III.3.6).

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

Considering a discretisation of $5,000,000$ points in time and $3,250$ points of space, we have obtained the following result. In particular, it is interesting because it has only H_0^1 regularity overall, and analytic regularity exactly outside of the control support: this is a pivot case where the reachability of the target is unknown for unconstrained controls (see [CR22]), let alone for the constrained case.

Proposition III.22. Considering the target

$$y_f : x \mapsto \begin{cases} \frac{1}{40}\frac{5x}{4} & \text{if } 0 \le x \le \frac{4}{5} \\ \frac{1}{40}(5 - 5x) & \text{if } \frac{4}{5} \le x \le 1. \end{cases} \tag{III.3.23}$$

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

$$J(p_f^{\text{reg}}) \in [-0.0017, -0.0009] < 0. \tag{III.3.24}$$

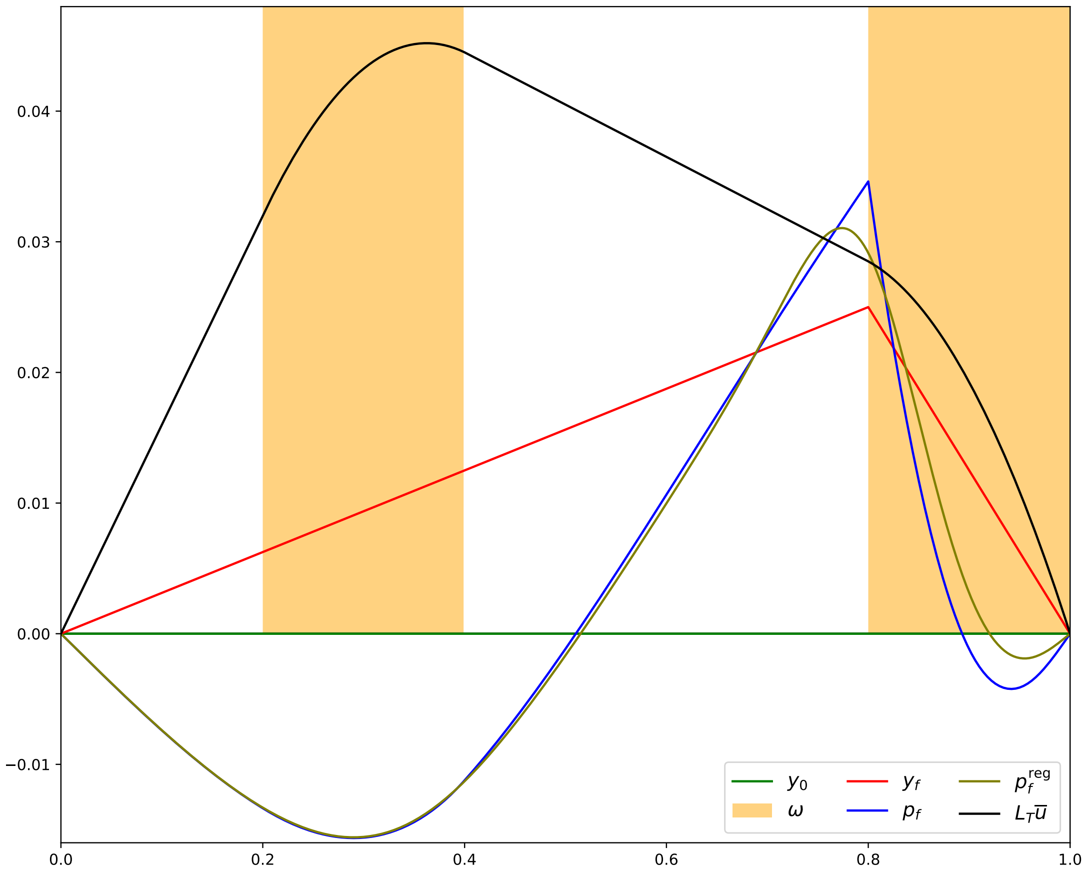


Figure III.3: Target and optimal dual certificates for Proposition III.22.

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

III.3.3.b Coupled 1D heat equation

In this section we shall consider the following dynamical system

$$\begin{cases} \partial_t y_1 - \kappa_1 \partial_{xx} y_1 = a y_1 + b y_2 & (t,x) \in (0,T] \times [0,1] \\ \partial_t y_2 - \kappa_2 \partial_{xx} y_2 = c y_1 + d y_2 + \chi_\omega u & (t,x) \in (0,T] \times [0,1], \\ y_i(0,x) = y_{0,i}(x) & i \in \{1,2\},\ x \in [0,1], \\ y_i(t,0) = y_i(t,1) = 0 & i \in \{1,2\},\ t \in (0,T]. \end{cases} \tag{III.3.25}$$

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

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

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

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

We consider the case where

$$\omega = (0, \tfrac{1}{2}), \quad \mathcal{U} := \{u \in U,\ -1 \leq u \leq 2\}. \tag{III.3.26}$$

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

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

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

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

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

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

$$\begin{aligned} \lambda_{\max}(S_1) \leq& \pi^2 \min(\kappa_1, \kappa_2) \\ \max\left(\lambda_{\max}(S_2), \lambda_{\max}(S_3)\right) \leq& \; 4\pi^2 \min(\kappa_1^2, \kappa_2^2). \end{aligned} \tag{III.3.27}$$

A possible choice of the continuity constant a_0 and the coercivity constant a_1 introduced in (III.2.2) are

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

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

$$C_0 = \frac{1}{2\pi\sqrt{2}}\left(\sqrt{4\pi^2 \min(\kappa_1^2, \kappa_2^2) - \max\left(\lambda_{\max}(S_2), 0\right)} + \sqrt{4\pi^2 \min(\kappa_1^2, \kappa_2^2) - \max\left(\lambda_{\max}(S_3), 0\right)}\right). \tag{III.3.28}$$

[a] In other words,

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

[b] In other words,

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

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

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

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

Computation of a_0. Let $x \in \mathcal{D}(A)$. According to the Cauchy-Schwarz and Poincaré inequalities, one has

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

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

$$\forall\, g \in H^2(0,1), \quad \inf_{v_h \in V_h} \|g - v_h\|_{H_0^1} \le \frac{h}{\sqrt{2}}\|g''\|_{L^2}, \tag{III.3.29}$$

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

$$\forall\, f \in X,\ \inf_{v_h \in V_h} \|A^{-1}f - v_h\|_V + \inf_{v_h \in V_h} \|(A^*)^{-1}f - v_h\|_V \le \frac{h}{\sqrt{2}} \left(\|\partial_{xx}(A^{-1}f)\| + \|\partial_{xx}((A^*)^{-1}f)\| \right). \tag{III.3.30}$$

It follows that it is enough to show

$$\forall\, f \in X, \quad \left(\|\partial_{xx}(A^{-1}f)\| + \|\partial_{xx}((A^*)^{-1}f)\| \right) \le C_0\sqrt{2}\|f\|. \tag{III.3.31}$$

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

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

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

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

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

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

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

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

$$\|Ag\|^2 \geq \left(\kappa_1^2 \|\partial_{xx} g_1\|_{L^2}^2 + \kappa_2^2 \|\partial_{xx} g_2\|_{L^2}^2\right) - \frac{1}{4\pi^2} \max\left(\lambda_{\max}(S_2), 0\right)(\|\partial_{xx} g_1\|_{L^2}^2 + \|\partial_{xx} g_2\|_{L^2}^2). \tag{III.3.32}$$

It follows that

$$\|Ag\| \geq \sqrt{\min(\kappa_1^2, \kappa_2^2) - \frac{1}{4\pi^2} \max\left(\lambda_{\max}(S_2), 0\right)}\, \|\partial_{xx} g\|. \tag{III.3.33}$$

In a similar manner, one can show that

$$\|A^* g\| \geq \sqrt{\min(\kappa_1^2, \kappa_2^2) - \frac{1}{4\pi^2} \max\left(\lambda_{\max}(S_3), 0\right)}\, \|\partial_{xx} g\|. \tag{III.3.34}$$

Applying (III.3.33) with $g = A^{-1} f$ and (III.3.34) with $g = (A^*)^{-1} f$ to (III.3.30) yields ($\mathcal{V}_1$) with

$$C_0 = \frac{1}{2\pi\sqrt{2}} \left(\sqrt{4\pi^2 \min(\kappa_1^2, \kappa_2^2) - \max\left(\lambda_{\max}(S_2), 0\right)} + \sqrt{4\pi^2 \min(\kappa_1^2, \kappa_2^2) - \max\left(\lambda_{\max}(S_3), 0\right)} \right). \tag{III.3.35}$$

□

For the following examples, we will use the constants

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ -2 & -1 \end{pmatrix}, \qquad \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1.5 \end{pmatrix}. \tag{III.3.36}$$

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

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

$$\mathcal{Y}_f = \{y \in X, \|y - y_f\| \leq \varepsilon\}. \tag{III.3.37}$$

One can then apply Remark III.3 with $M_{\mathcal{Y}_f} = \|y_f\| + \varepsilon$ and use the same method to try and prove the non-reachability of all elements of $\mathcal{Y}_f$ under the constraints III.3.26. In this context, one has that

$$\forall p_f \in X, \quad \sigma_{\mathcal{Y}_f}(p_f) = \langle y_f, p_f \rangle + \varepsilon \|p_f\|. \tag{III.3.38}$$

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

Proposition III.24. Let

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

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

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

Notice that once more, the dual certificate $\binom{p_1}{p_2}$ tends to resemble $\mathcal{Y}_f + \{-S_T y_0\}$ so as to minimise $\sigma_{\mathcal{Y}_f}(-p_f) + \langle S_T y_0, p_f \rangle$, while staying close to 0 on ω, to minimise $\int_0^T \sigma_{B\mathcal{U}}(S_t^* p_f)\,\mathrm{d}t$. Notice as well that the discretisation to prove this result is much coarser: yet, the turnpoint between discretisation errors (here, $8.1 \cdot 10^{-3}$) and round-off errors ($4.2 \cdot 10^{-3}$) is close. This

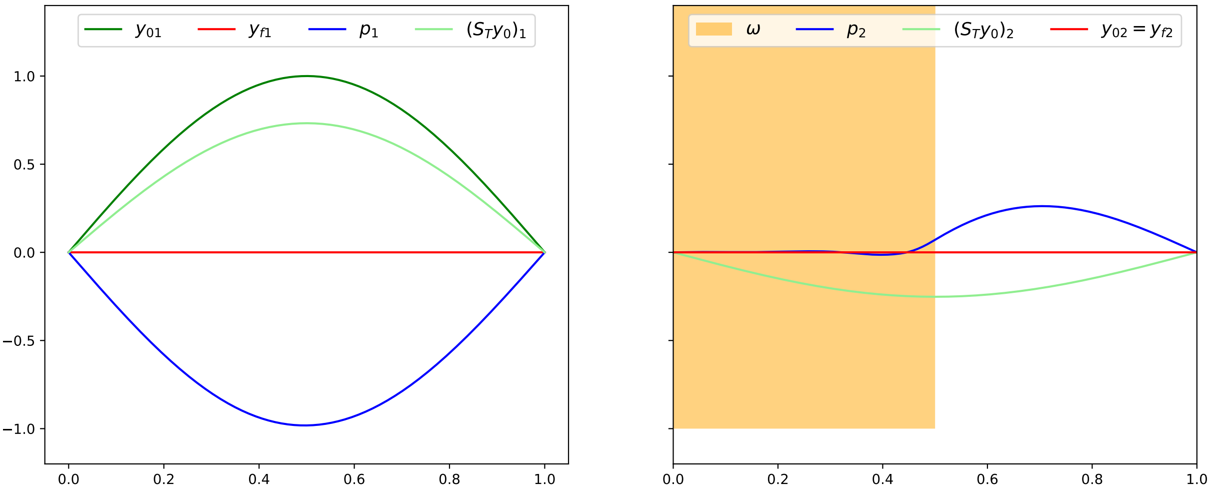


Figure III.4: Initial and final state without control, target, control domain and dual certificate for Proposition III.24.

is mainly due on the one hand to the additional term of $S_T^* p_f$ which considerably increases discretisation errors, and on the other hand that the space-discretisation has doubled complexity with respect to the single equation case: two intervals $[0,1]$ are discretised with 750 points, increasing rounding errors.

> **Remark III.25.** In the spirit of Lemma III.6, since $\ker(B) \cap \mathcal{U} \neq \emptyset$ and for all $y \in X$ and $t > 0$, $\|S_t y\| \leq 1$, one can show that such a T is a certified lower-bound of the minimal time $T^\star(y_0, \mathcal{Y}_f)$ needed to reach $\mathcal{Y}_f$ from y_0, where "minimal" means that for any time $t > T^\star(y_0, \mathcal{Y}_f)$, the target set $\mathcal{Y}_f$ is $\mathcal{U}$-reachable from y_0 in time t.

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

$$\mathcal{Y}_f := \{y_1\} \times L^2(0,1). \tag{III.3.39}$$

In this context, we find

$$\forall\, p_f = \begin{pmatrix} p_f^1 \\ p_f^2 \end{pmatrix} \in X, \quad \sigma_{\mathcal{Y}_f}(p_f) = \begin{cases} \langle y_1, p_f^1 \rangle & \text{if } p_f^2 = 0 \\ +\infty & \text{if } p_f^2 \neq 0. \end{cases} \tag{III.3.40}$$

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

$$\forall\, p_f \in X_1, \quad J(p_f; \mathcal{Y}_f) = \int_0^T \sigma_{\mathcal{U}}(L_T^* p_f(t))\,\mathrm{d}t - \langle y_f, p_f \rangle, \tag{III.3.41}$$

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

> **Proposition III.26.** Let
>
> $$y_1 = -\frac{1}{50}\sin(2\pi\cdot) \quad \text{and} \quad \mathcal{Y}_f := \{y_1\} \times L^2(0,1).$$
>
> $\mathcal{Y}_f$ is not reachable in time $T = 1$. Indeed, for $p_f = \left(\begin{smallmatrix} y_1 \\ 0 \end{smallmatrix}\right)$, we have that
>
> $$J(p_f; \mathcal{Y}_f) \in [-0.0262, -0.0008] < 0.$$

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

III.4 Appendix

III.4.1 Complex analysis

Lemma III.27 (L. Thomassey)**.** Let $z \in \mathbb{C}$ such that $\mathrm{Re}\,(z) \leq 0$. Then

$$\left| e^z - \frac{1}{1-z} \right| \leq \frac{1}{2}|z|^2. \tag{III.4.1}$$

Proof. Let $z \in \mathbb{C}$ such that $\mathrm{Re}\,(z) \leq 0$. First,

$$\left| e^z - \frac{1}{1-z} \right| \leq |e^z| + \frac{1}{|1-z|} \leq 1 + 1 = 2. \tag{III.4.2}$$

In particular, since $2 \leq \frac{1}{2}|z|^2$ for all $|z| \geq 2$, the inequality is proved for z such that $|z| \geq 2$. Let us now suppose that $z = iy, y \in \mathbb{R}$. Then

$$\begin{aligned}
\left| e^z - \frac{1}{1-z} \right| \leq \frac{1}{2}|z|^2 &\iff \left| e^{iy} - \frac{1}{1-iy} \right| \leq \frac{1}{2}|iy|^2 \\
&\iff |e^{iy}(1-iy) - 1| \leq \frac{1}{2}y^2|1-iy| \\
&\iff (\cos(y) + y\sin(y) - 1)^2 + (\sin(y) - y\cos(y))^2 \leq \frac{1}{4}y^4(1+y^2) \\
&\iff 2 + y^2 - 2y\sin(y) - 2\cos(y) \leq \frac{1}{4}y^4(1+y^2)
\end{aligned}$$

Let us hence prove the final inequality, by letting $f : x \mapsto 2 + x^2 - 2x\sin(x) - 2\cos(x)$. f, which has derivative $f' : x \mapsto= 2x(1-\cos(x))$ for all $x \in \mathbb{R}$. Using the convex inequality $1 - \cos(y) \leq \frac{1}{2}y^2$ we deduce that $f'(y) \leq y^3$ for all $y \geq 0$.

Hence, since $f(0) = 0$, we infer

$$f(y) = f(0) + \int_0^y f(t)\,\mathrm{d}t \leq \int_0^y t^3\,\mathrm{d}t = \frac{1}{4}y^4. \tag{III.4.3}$$

Since f is even, we conclude that for all $y \in \mathbb{R}$,

$$2 + y^2 - 2y\sin(y) - 2\cos(y) \leq \frac{1}{4}y^4 \leq \frac{1}{4}y^4(1+y^2), \tag{III.4.4}$$

as wanted. We are left with proving the main result for $z \in A := \{z \in \mathbb{C},\ \mathrm{Re}\,(z) < 0, |z| < 2\}$. Remark that

$$A \subset B := \{z \in \mathbb{C},\ \max(|\mathrm{Re}\,(z)|, |\,\mathrm{Im}(z)|) \leq 2,\ \mathrm{Re}\,(z) \leq 0\}. \tag{III.4.5}$$

Let $g : z \mapsto \frac{1-z}{z^2}(e^z - \frac{1}{1-z}) = \frac{1}{z^2}(e^z(1-z) - 1)$, which is a holomorphic function on $\mathbb{C}\backslash\{0\}$. Furthermore, the singularity of g at 0 is clearly removable. Applying the complex maximum principle to (the holomorphic extension of) g on B, we obtain

$$\sup_{z \in B} |g(z)| = \sup_{z \in \partial B} |g(z)|, \tag{III.4.6}$$

and thus $|g(z)| \leq \frac{1}{2}$, for all $z \in A \subset B$, which ends the proof. □

III.4.2 Functional analysis

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

Throughout this section, we assume that we are in the setting of Section III.2, with an operator $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$ and discretisation subspaces V_h satisfying (III.2.2) and $(\mathcal{V}_1)$.

Let α and β satisfy $0 \leq \alpha < \alpha + \beta < \frac{\pi}{2}$. Recall the sectors defined by equations (III.2.3) and (III.2.4),and thus that $\mathcal{A}$ is a mα-accretive operator. A classical consequence of the Lumer-Phillips theorem (see Theorem I.13) is that $-\mathcal{A}$ generates of a semigroup of contraction, and thus for all $t \in \mathcal{S}_\beta$, the function $S_t := \exp(-t\mathcal{A})$ is well defined.

Theorem III.28. For all $t \in \mathcal{S}_\beta$, with $t \neq 0$ and $0 \leq \alpha < \alpha + \beta < \frac{\pi}{2}$, and for every integer $k \geq 0$, the operator $S_t \in \mathcal{L}(X, \mathcal{D}(\mathcal{A}^k))$. Moreover, the map $t \mapsto S_t$, from $\mathcal{S}_{\frac{\pi}{2}-\alpha}$ into $\mathcal{L}(X)$, is holomorphic on $\operatorname{Int} \mathcal{S}_{\frac{\pi}{2}-\alpha}$, and we have the estimate:

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

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

$$S_t = (\operatorname{Id} + \mathcal{A})^{-k} f_k(\mathcal{A}) \in \mathcal{L}(X, \mathcal{D}(\mathcal{A}^k)). \tag{III.4.7}$$

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

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

Using Theorem I.7:

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

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

$$\forall\, k \in \mathbb{N}, \quad S_t^{(k)} = (-\mathcal{A})^k S_t. \tag{III.4.8}$$

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

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

Cauchy's residue theorem gives:

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

leading to:

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

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

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

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

Theorem III.29. Given $z_0 \in X$, the unique solution $z \in \mathcal{C}^1((0,\infty);\mathcal{D}(\mathcal{A})) \cap \mathcal{C}^0([0,\infty);X)$ to

$$\begin{cases} z'(t) = -\mathcal{A}z(t) \\ z(0) = z_0 \end{cases} \tag{III.4.9}$$

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

$$\forall\, k,\ell \geq 0, \forall\, t > 0, \quad \|z^{(k+l)}(t)\| = \|\mathcal{A}^k z^{(l)}(t)\| \leq \frac{k!}{(t\cos(\alpha))^k}\|\mathcal{A}^\ell z_0\|. \tag{III.4.10}$$

Proof. It is clear that $z(t)$ is a solution of (P) since $S_t' + \mathcal{A}S_t = 0$. The C^∞ regularity on $(0,\infty)$ follows from the previous theorem III.28, while continuity at $t=0$ in X results from the strong continuity of the semigroup. Since the problem is linear, to prove uniqueness of the solution to (III.4.9), it suffices to show that $z_0 = 0$ implies $z(t) = 0$. Indeed, if z is a solution of (III.4.9) and $z_0 = 0$, then:

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

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

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

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

and by induction:

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

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

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

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

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

We then observe that $z^{(\ell)}$ is a solution of problem (III.4.9) with z_0 replaced by $(-\mathcal{A})^\ell z_0$, from which it follows that $z^{(\ell)} \in C^0([0,\infty);X)$. The estimate (III.4.10) then results as a consequence of Theorem III.28. □

Proposition III.30 (Time approximation)**.** For $z_0 \in \mathcal{D}(\mathcal{A})$, we consider

$$\begin{cases} \partial_t z = -\mathcal{A}z \\ z(0) = z_0. \end{cases} \tag{P}$$

Define $N_0 \in \mathbb{N}^*$, $\Delta t = \frac{T}{N_0}$, and for $n \in \{0,\dots,N_0\}$, $t_n = n\Delta t$. Discretise the system using the implicit Euler scheme:

$$\begin{cases} z_0 = z_0 \\ (\mathrm{Id} + \Delta t\mathcal{A})z_{n+1} = z_n \quad \forall\, n \in \{0,\dots,N_0-1\}. \end{cases} \tag{III.4.11}$$

Then

$$\forall\, n \in \{0,\dots,N_0\}, \quad \|z(t_n) - z_n\| \leq \Delta t\,\frac{C_\alpha}{\cos\alpha}\|\mathcal{A}z_0\|. \tag{III.4.12}$$

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

Proof. By definition, for all $n \in \{0, \dots, N_0\}$, we have

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

Hence me may write

$$z(t_n) - z_n = \Delta t\, \varphi_n(\Delta t\mathcal{A})\, \mathcal{A}z_0, \tag{III.4.13}$$

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

$$\varphi_n(z) = \frac{e^{-nz} - (1+z)^{-n}}{z}, \tag{III.4.14}$$

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

Using Theorem III.8, we find

$$\|z(t_n) - z_n\| \leq \Delta t\, \|\varphi_n(\Delta t\mathcal{A})\|_{\mathcal{L}(X)}\, \|\mathcal{A}z_0\| \leq \Delta t\, C_\alpha \sup_{z\in\mathcal{S}_\alpha} |\varphi_n(z)|\, \|\mathcal{A}z_0\|. \tag{III.4.15}$$

Let us finish with the proof that $\sup_{z\in\mathcal{S}_\alpha} |\varphi_n(z)| \leq \frac{1}{\cos(\alpha)}$. Let $z \in \mathcal{S}_\alpha$, $z \neq 0$ (the case $z = 0$ is obvious). Then $\operatorname{Re}(z) > 0$, hence by Lemma III.27 applied to $-z$,

$$|\varphi_1(z)| = \frac{1}{|z|}\left[e^{-z} - \frac{1}{1+z}\right] \leq \frac{1}{2}|z|. \tag{III.4.16}$$

Since

$$\varphi_n(z) = \frac{1}{z}\left[e^{-z} - \frac{1}{1+z}\right] \sum_{k=0}^{n-1} e^{-kz}\left(\frac{1}{1+z}\right)^{n-k-1}, \tag{III.4.17}$$

we get

$$|\varphi_n(z)| \leq \frac{1}{2}|z| \sum_{k=0}^{n-1} \left|e^{-z}\right|^k \left|\frac{1}{1+z}\right|^{n-k-1}. \tag{III.4.18}$$

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

$$|1+z| = |1+\rho e^{i\theta}| \geq |1+\rho e^{i\alpha}| \geq 1 + \rho\cos(\alpha). \tag{III.4.19}$$

Thus

$$\left|\frac{1}{1+z}\right| \leq \frac{1}{1+\rho\cos(\alpha)}. \tag{III.4.20}$$

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

$$|\varphi_n(z)| \leq \frac{1}{2}\rho \sum_{k=0}^{n-1} \left(e^{-\rho\cos(\alpha)}\right)^k \left(\frac{1}{1+\rho\cos(\alpha)}\right)^{n-k-1}. \tag{III.4.21}$$

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

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

Finally

$$|\varphi_n(z)| \leq \frac{1}{2}\rho n \frac{1}{(1+\rho\cos(\alpha))^{n-1}}, \tag{III.4.22}$$

and maximising the right-hand side for $\rho > 0$ yields the desired result if $n \geq 2$. As for $n = 1$, if $|z| \leq 2$ the upper bound from Lemma III.27 proves that $|\varphi_1(z)| \leq 1 \leq \frac{1}{\cos(\alpha)}$, and similarly and if $|z| \geq 2$, $|\varphi_1(z)| \leq \frac{2}{|z|} \leq 1 \leq \frac{1}{\cos(\alpha)}$. □

Proposition III.31 (Spatial approximation)**.** Let $z_{h,0} \in V_h$, $z_h : t \mapsto S_{t,h}z_{h,0}$, where

$S_{t,h} = \exp(-t\mathcal{A}_h)$. For $z_0 \in \mathcal{D}(\mathcal{A})$, $z(t) = \exp(-t\mathcal{A})z_0$, there holds

$$\forall\, t > 0, \quad \|z(t) - z_h(t)\| \leq \|z_0 - z_{h,0}\| + \left(6 + \frac{4\ln(2)}{\cos(\alpha)}\right) C_1 h^2 \|\mathcal{A}z_0\|. \tag{III.4.23}$$

Proof. For $t > 0$, we write

$$\begin{aligned}\|z(t) - P_h z(t)\| &\leq \|z(t) - \mathcal{A}_h^{-1} P_h \mathcal{A} z(t)\| \text{ since } \mathcal{A}_h^{-1} P_h \mathcal{A} z(t) \in V_h \\ &= \|(\mathcal{A}^{-1} - \mathcal{A}_h^{-1} P_h)(\mathcal{A}z(t))\| \\ &\leq C_1 h^2 \|\mathcal{A}z(t)\| \text{ using } (\mathcal{H}_1).\end{aligned} \tag{III.4.24}$$

Similarly,

$$\begin{aligned}\|z'(t) - P_h z'(t)\| \quad &\leq \|z'(t) - \mathcal{A}_h^{-1} P_h \mathcal{A} z'(t)\| \leq C_1 h^2 \|\mathcal{A}z'(t)\| \\ &\leq \frac{C_1 h^2}{t\cos(\alpha)} \|\mathcal{A}z_0\| \text{ using Theorem III.29.}\end{aligned} \tag{III.4.25}$$

Using the triangle inequality, we deduce the upper bounds

$$\|(P_h - \mathcal{A}_h^{-1} P_h \mathcal{A}) z(t)\| \leq 2C_1 h^2 \|\mathcal{A}z(t)\| \quad \text{and} \quad \|(P_h - \mathcal{A}_h^{-1} P_h \mathcal{A}) z'(t)\| \leq 2\frac{C_1 h^2}{t\cos(\alpha)} \|\mathcal{A}z_0\|. \tag{III.4.26}$$

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

$$\|z(t) - z_h(t)\| \leq \|e_h(t)\| + \|z(t) - \mathcal{A}_h^{-1} P_h \mathcal{A} z(t)\| \leq \|e_h(t)\| + C_1 h^2 \|\mathcal{A}z(t)\|. \tag{III.4.27}$$

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

$$e_h'(t) + \mathcal{A}_h e_h(t) = \mathcal{A}_h^{-1} P_h \mathcal{A} z'(t) + P_h \mathcal{A} z(t) = (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(t). \tag{III.4.28}$$

Duhamel's formula yields

$$e_h(t) = S_{t,h} e_h(0) + \int_0^t S_{t-\sigma,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(\sigma) \,\mathrm{d}\sigma =: E_1 + E_2 + E_3, \tag{III.4.29}$$

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

First, we have, using $\|S_{t,h}\|_{\mathcal{L}(X)} \leq 1$ and $(\mathcal{H}_1)$

$$\|E_1\| \leq \|e_h(0)\| = \|z_{h,0} - \mathcal{A}_h^{-1} P_h \mathcal{A} z_0\| \leq \|z_0 - z_{h,0}\| + \|(\mathcal{A}^{-1} - \mathcal{A}_h^{-1} P_h)(\mathcal{A}z_0)\| \tag{III.4.30}$$

$$\leq \|z_0 - z_{h,0}\| + C_1 h^2 \|\mathcal{A}z_0\|. \tag{III.4.31}$$

Secondly,

$$\begin{aligned}\|E_2\| &\leq \int_{t/2}^t \|S_{t-\sigma,h} (\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(\sigma)\| \,\mathrm{d}\sigma \leq \int_{t/2}^t \|(\mathcal{A}_h^{-1} P_h \mathcal{A} - P_h) z'(\sigma)\| \,\mathrm{d}\sigma \\ &\leq \int_{t/2}^t 2\frac{C_1 h^2}{\sigma\cos(\alpha)} \|\mathcal{A}z_0\| \,\mathrm{d}\sigma = \frac{2\ln(2)}{\cos(\alpha)} C_1 h^2 \|\mathcal{A}z_0\|,\end{aligned} \tag{III.4.32}$$

where we used (III.4.26).

For the last term E_3, we first integrate by parts, then use (III.4.26) to uncover

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

It follows from Theorem III.29 that

$$\|\mathcal{A}_h S_{t-\sigma,h}\|_{\mathcal{L}(X)} \leq \frac{1}{(t-\sigma)\cos(\alpha)}, \tag{III.4.33}$$

hence

$$\|E_3\| \leq \left(4 + \frac{2\ln(2)}{\cos(\alpha)}\right) C_1 h^2 \|\mathcal{A} z_0\|. \tag{III.4.34}$$

Gathering all the estimates, we obtain

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

□

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

$$\begin{aligned} &\|z(t_n) - z_{h,n}\| \leq E_1 + E_2 + E_3 \\ &\text{with } \begin{cases} E_1 = \|z(t_n) - S_{t_n,h}\mathcal{A}_h^{-1} P_h \mathcal{A} z_0\| \\ E_2 = \|(S_{t_n,h} - (\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n})\mathcal{A}_h^{-1} P_h \mathcal{A} z_0\| \\ E_3 = \|(\mathrm{Id} + \Delta t \mathcal{A}_h)^{-n}(\mathcal{A}_h^{-1} P_h \mathcal{A} z_0 - z_{h,0})\| \end{cases} \end{aligned} \tag{III.4.35}$$

It follows from Proposition III.31 with $z_{h,0} = \mathcal{A}_h^{-1} P_h \mathcal{A} z_0$ and $(\mathcal{H}_1)$ that

$$E_1 \leq \|z_0 - \mathcal{A}_h^{-1} P_h \mathcal{A} z_0\| + \left(6 + \frac{4\ln(2)}{\cos(\alpha)}\right) C_1 h^2 \|\mathcal{A} z_0\| \leq \left(7 + \frac{4\ln(2)}{\cos(\alpha)}\right) C_1 h^2 \|\mathcal{A} z_0\|. \tag{III.4.36}$$

Using Proposition III.30 with $\mathcal{A}$ replaced by $\mathcal{A}_h$,

$$E_2 \leq \frac{\Delta t}{\cos(\alpha)} C_\alpha \|\mathcal{A}_h \, \mathcal{A}_h^{-1} P_h \mathcal{A} z_0\| \leq \frac{\Delta t}{\cos(\alpha)} C_\alpha \|\mathcal{A} z_0\|. \tag{III.4.37}$$

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

$$\begin{aligned} E_3 &\leq C_\alpha \|\mathcal{A}_h^{-1} P_h \mathcal{A} z_0 - z_{h,0}\| \\ &\leq C_\alpha (\|\mathcal{A}_h^{-1} P_h \mathcal{A} z_0 - z_0\| + \|z_0 - z_{h,0}\|) \\ &\leq C_\alpha C_1 h^2 \|\mathcal{A} z_0\| + C_\alpha \|z_0 - z_{h,0}\|. \end{aligned} \tag{III.4.38}$$

Finally,

$$\|z(t_n) - z_{h,n}\| \leq C_\alpha \|z_0 - z_{h,0}\| + \left(C_1 \left(7 + \frac{4\ln(2)}{\cos(\alpha)} + C_\alpha\right) h^2 + \frac{C_\alpha}{\cos(\alpha)} \Delta t\right) \|\mathcal{A} z_0\|. \tag{III.4.39}$$

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

$$C_2 = C_1 \left(7 + 4\ln(2)\frac{a_1}{a_0} + C_\alpha\right), \qquad C_3 = \frac{a_1}{a_0} C_\alpha. \tag{III.4.40}$$

□

IV

Certified reachability for finite-dimensional control problems

Contents

Abstract

In this chapter, we present current research projects on computer-assisted proofs of reachability: given an unknown compact convex reachable set $\mathcal{R} \subset \mathbb{R}^n$ for which the support function $\sigma_{\mathcal{R}}$ can be computed, we develop under- and over-approximations of $\mathcal{R}$ using polytopes. We outline research directions to either iteratively compute guaranteed approximations of the reachable set, or to focus on the reachability of a single target.

IV.1 Introduction

In this chapter, we present ongoing work jointly undertaken with Maxime Breden, Camille Pouchol, Yannick Privat and Christophe Zhang. The ultimate goal of these efforts is to provide computer-assisted proofs of reachability for finite-dimensional linear control systems under bounded control constraints; consider the system

$$\begin{cases} y'(t) = Ay(t) + Bu(t) & \text{for a.e. } t \in [0,T], \\ y(0) = y_0 \in \mathbb{R}^n, \\ u(t) \in \mathcal{U} & \text{for a.e. } t \in [0,T], \end{cases} \tag{$\mathcal{S}$}$$

where $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, $B \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ are two matrices, and $\mathcal{U} \subset \mathbb{R}^m$ is a compact set. The solution is then characterised as

$$y : t \mapsto y(t) = e^{tA} y_0 + L_t u. \tag{IV.1.1}$$

Considering a convex and compact initial set $\mathcal{Y}_0 \subset \mathbb{R}^n$ and a closed convex target set $\mathcal{Y}_f$, recall that $\mathcal{Y}_f$ is said to be $\mathcal{U}$-reachable from $\mathcal{Y}_0$ in time T if

$$L_T E_{\mathcal{U}} \cap (\mathcal{Y}_f - e^{TA} \mathcal{Y}_0) \neq \emptyset, \tag{IV.1.2}$$

where $E_{\mathcal{U}} = \{u \in L^2(0, T; \mathbb{R}^m), \quad \text{for a.e. } \; t \in [0, T], u(t) \in \mathcal{U}\}$. The reachability property was proved to be equivalent to the global nonnegativity of the separating functional

$$J : \begin{cases} \mathbb{R}^n & \to \mathbb{R} \\ p_f & \mapsto \sigma_{L_T E_{\mathcal{U}}}(p_f) + \sigma_{\mathcal{Y}_f - e^{TA}\mathcal{Y}_0}(-p_f); \end{cases} \tag{IV.1.3}$$

which can be interpreted as J quantifying the separation between $L_T E_{\mathcal{U}} - \mathcal{Y}_f + e^{TA}\mathcal{Y}_0$ and 0, leading to the following reformulation:

$$\forall p_f \in \mathbb{R}^n, \quad J(p_f) = \sigma_{e^{TA}\mathcal{Y}_0 + L_T E_{\mathcal{U}} - \mathcal{Y}_f}(p_f). \tag{IV.1.4}$$

In this chapter, we explore two main objectives:

- firstly, as was seen before, the reachability of $\mathcal{Y}_f$ is equivalent to the overall nonnegativity of J. We will therefore propose different options to prove the nonnegativity of J.
- A more challenging problem is that of the approximation of the reachable set: since it is closed and convex, this is equivalent to the characterisation of its support function $\sigma_{e^{TA}\mathcal{Y}_0 + L_T E_{\mathcal{U}}}$.

Both objectives will be explored under various assumptions. Note that this is a topic of ongoing research, so actual results are scarce and most of the chapter presents intuitions as to how one might tackle these objectives. Although computer-assisted proofs of reachability or rigorous under-approximations (contained in the reachable set) and over-approximations (which contain it) of the reachable set are the long-term goal of this chapter, most of this chapter is devoted to theoretical studies, where no error is considered.

Note that approximations of the reachable set have been extensively studied in the literature, including from the viewpoint of support functions: see Section 0.3 for a detailed account of the proposed methods. Certified over-approximations have also been considered [Imm15], but to our knowledge no certified under-approximations have been proposed yet. Also, we are not aware of computer-assisted methods focusing on proofs of reachability for a single target set.

In light of the theoretical purpose of this chapter, we shall place ourselves in a purely convex analytic framework, considering an unknown nonempty compact convex set $\mathcal{R} \subset \mathbb{R}^n$, for which we shall only assume knowledge of its support function. $\mathcal{R}$ might thus correspond to the reachable set $e^{TA}\mathcal{Y}_0 + L_T E_{\mathcal{U}}$ or to the reachable set recentred around a given target $e^{TA}\mathcal{Y}_0 + L_T E_{\mathcal{U}} - \mathcal{Y}_f$.

Among the possible approximation techniques of $\mathcal{R}$ that can be explored, two are particularly interesting.

> **Definition IV.1.** Let $\mathcal{R} \subset \mathbb{R}^n$. An over-approximation (or outer approximation) of $\mathcal{R}$ is a set $\mathcal{O} \subset \mathbb{R}^n$ such that $\mathcal{R} \subset \mathcal{O}$, and an under-approximation (or inner approximation) of $\mathcal{R}$ is a set $\mathcal{I} \subset \mathbb{R}^n$ such that $\mathcal{I} \subset \mathcal{R}$.

Over-approximations are relevant in order to prove the non-reachability of a target, while under-approximations provide proofs of reachability. In this chapter, we will study polytopal over- and under-approximations of nonempty compact convex sets. We refer to Subsection I.2.5 for the most useful standard definitions and properties of convex sets and polytopes.

This chapter will be divided as follows:

- Section IV.2 focuses on computing over-approximations of $\mathcal{R}$ and provides an interpretation using polar polytopes;
- Section IV.3 introduces methods for computing under-approximations of $\mathcal{R}$ based solely on evaluations of its support function, as well as over-approximations of the polar $\mathcal{R}^\circ$;
- Section IV.4 uses the additional knowledge of the maximisers of the support function to provide more accurate under-approximation methods;
- Section IV.5 discusses iterative methods for constructing under- and over-approximations of $\mathcal{R}$, both to efficiently approximate $\mathcal{R}$, or to determine whether $0 \in \mathcal{R}$;
- Finally, Section IV.6 discusses the introduction of errors (both discretisation and round-off) on the support function and how these might affect the different methods discussed in the earlier sections.

We advise the reader that most of what will be introduced is not new with regard to the literature. Indeed, most of this chapter is concerned with polytopal approximations of convex sets, which has been an extensively studied topic (see for example the survey [Bro08]). In particular, Sections IV.2, IV.4 and IV.5 do not present new results. In contrast, to our knowledge the under-approximation presented in Section IV.3 has not been studied, nor have the rigorous approximations discussed in Section IV.6.

Overall, this chapter is the basis of ongoing work aiming to certify these approximations using computer-assisted proofs. We hope it will provide a first framework allowing one to build some intuition – which is why many figures have been added – and perhaps lead to new computer-assisted proofs of reachability.

IV.2 Polytopal over-approximations of a compact convex set

In this section, we are concerned with over-approximations, and present one classical way to produce a polytopal over-approximation of a nonempty compact convex set $\mathcal{R}$. Let us first recall a classical result, the proof of which can be found in [Sch13, Theorem 1.8.19]:

Theorem IV.2. Let $\mathcal{R} \subset \mathbb{R}^n$ a convex set containing 0. Then for all $\lambda > 1$, there exists a polytope P such that

$$P \subset \mathcal{R} \subset \lambda P. \tag{IV.2.1}$$

This theorem in itself justifies that polytopal approximations of a convex set $\mathcal{R}$ can converge towards $\mathcal{R}$. We shall now provide a method to compute polytopal over-approximations.

IV.2.1 Polytopal over-approximation of $\mathcal{R}$

Recall the definition of the support function:

$$\forall\, p \in \mathbb{R}^n, \quad \sigma_{\mathcal{R}}(p) = \sup_{x \in \mathcal{R}} \langle p, x \rangle. \tag{IV.2.2}$$

Therefore, for any $p \in \mathbb{R}^n$,

$$\mathcal{R} \subset \left\{ x \in \mathbb{R}^n,\ \langle p, x \rangle \leq \sigma_{\mathcal{R}}(p) \right\}, \tag{IV.2.3}$$

which allows us to build over-approximations of $\mathcal{R}$, using the following proposition, whose last equivalence is a direct consequence of Proposition I.43.

Proposition IV.3. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set, and let $(p_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m$. Defining

$$\forall\, j \in \{1,\dots,m\}, \quad \sigma_{\mathcal{R}}(p_j) \leq b_j \in \mathbb{R}, \tag{IV.2.4}$$

we have that

$$\mathcal{R} \subset \bigcap_{j=1}^{m} \{x \in \mathbb{R}^n,\ \langle p_j, x\rangle \leq b_j\} =: P. \tag{IV.2.5}$$

Furthermore, the over-approximation P is a bounded polytope if and only if $0 \in \operatorname{Int}(\operatorname{conv}((p_j)_j))$.

Therefore, one can easily bound $\mathcal{R}$ from the outside: Figure IV.1(a) presents one example of $\mathcal{R}$ and its polytope approximation using five directions, that is, created with five supporting hyperplanes of $\mathcal{R}$.

Let us introduce a convenient definition characterising polytopal approximations.

Definition IV.4. Given a nonempty compact convex set $\mathcal{R} \subset \mathbb{R}^n$. Let $P \subset \mathbb{R}^n$ be a polytope defined by

$$P = \bigcap_{j=1}^{m} \{x \in \mathbb{R}^n,\ \langle p_j, x\rangle \leq b_j\}. \tag{IV.2.6}$$

We say that P is a sharp polytopal over-approximation of $\mathcal{R}$ if $\mathcal{R} \subset P$ and if furthermore

$$\forall\, j \in \{1,\dots,k\}, \quad \sigma_{\mathcal{R}}(p_j) = b_j. \tag{IV.2.7}$$

Similarly, if P is defined by

$$P = \operatorname{conv}((y_i)_{i\in\{1,\dots,k\}}), \tag{IV.2.8}$$

we say that P is a sharp polytopal under-approximation of $\mathcal{R}$ if $P \subset \mathcal{R}$ and if furthermore

$$\forall\, i \in \{1,\dots,k\}, \exists\, q_i \in \mathbb{R}^n, \quad \sigma_{\mathcal{R}}(q_i) = \langle q_i, y_i\rangle. \tag{IV.2.9}$$

Since $\mathcal{R}$ is compact, this definition guarantees that if P is a sharp over-approximation, $\mathcal{R}$ touches each facet of P, and similarly if P is a sharp under-approximation, each vertex of P lies on the boundary of $\mathcal{R}$. By definition, the polytopal approximations introduced in Proposition IV.3 are sharp over-approximations of $\mathcal{R}$ if and only if $\forall\, j,\ b_j = \sigma_{\mathcal{R}}(p_j)$.

Remark as well that for a given set of $(p_j)_j$ or $(y_i)_i$, there exists exactly one sharp polytopal over-approximation of a convex set $\mathcal{R}$ defined by $(p_j)_j$, and similarly there exists exactly one sharp polytopal under-approximation of $\mathcal{R}$ defined by $(y_i)_i$. Therefore, when defining sharp polytopal under- and over-approximations, we shall only say that it is defined using $\mathcal{R}$ and either $(p_j)_j$ or $(y_i)_i$.

As was seen in Section I.2.5, a useful tool to study convex bodies is their polar set. The following section links over-approximations of $\mathcal{R}$ with under-approximations of its polar $\mathcal{R}^\circ$.

IV.2.2 Polytopal under-approximation of $\mathcal{R}^\circ$

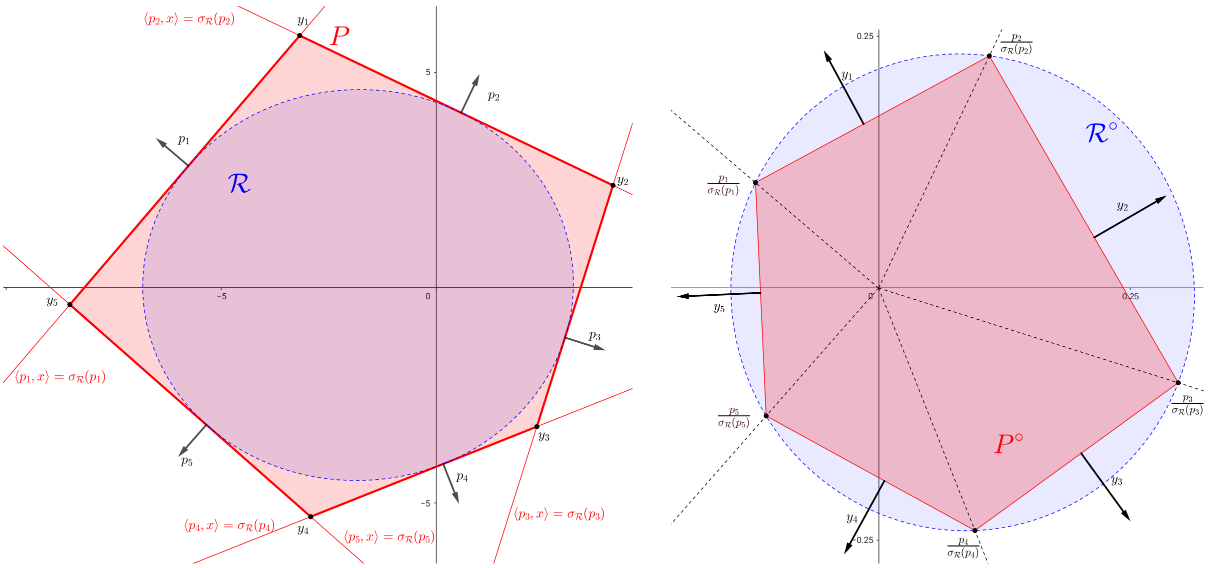
As was seen in Section I.2.5, constructing an over-approximation of $\mathcal{R}$ is equivalent to constructing an under-approximation of $\mathcal{R}^\circ$. Indeed,

$$\mathcal{R} \subset P \quad \Longleftrightarrow \quad P^\circ \subset \mathcal{R}^\circ. \tag{IV.2.10}$$

Using Proposition I.37, we have the following more precise statement.

Proposition IV.5. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set. Let $(p_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m$ define a sharp polytopal over-approximation of $\mathcal{R}$. Then, if $0 \in \operatorname{Int}(P)$, P° is a sharp polytopal under-approximation of $\mathcal{R}^\circ$, and

$$P^\circ = \operatorname{conv}\left(\left(\frac{p_j}{\sigma_{\mathcal{R}}(p_j)}\right)_j\right). \tag{IV.2.11}$$



$\langle p_2, x\rangle = \sigma_{\mathcal{R}}(p_2)$
y_1
P
5
p_2
p_1
$\mathcal{R}$
y_2
y_5
-5
0
$\langle p_1, x\rangle = \sigma_{\mathcal{R}}(p_1)$
p_3
y_3
p_5
p_4
-5
$\langle p_4, x\rangle = \sigma_{\mathcal{R}}(p_4)$
y_4
$\langle p_5, x\rangle = \sigma_{\mathcal{R}}(p_5)$
$\langle p_3, x\rangle = \sigma_{\mathcal{R}}(p_3)$
0.25
$\frac{p_2}{\sigma_{\mathcal{R}}(p_2)}$
y_1
$\mathcal{R}^\circ$
$\frac{p_1}{\sigma_{\mathcal{R}}(p_1)}$
y_2
0
0.25
y_5
$\frac{p_3}{\sigma_{\mathcal{R}}(p_3)}$
$\frac{p_5}{\sigma_{\mathcal{R}}(p_5)}$
P°
y_3
y_4
-0.25
$\frac{p_4}{\sigma_{\mathcal{R}}(p_4)}$

Figure IV.1: Example of polytopal over-approximation (left panel (a)) and its dual under-approximation (right panel (b)).

Figure IV.1 presents both approximations of $\mathcal{R}$ using P and $\mathcal{R}^\circ$ using P°: while $\mathcal{R}$ is circumscribed to P, P° is evidently inscribed in $\mathcal{R}^\circ$. One can clearly see what would allow the evaluation of J at another direction: in the primal setting, it would correspond to adding a facet to P touching $\mathcal{R}$, whereas for the polar, it would add a vertex of P° on the boundary of $\mathcal{R}^\circ$.

Polar sets are useful tools when studying polytopes because they allow for a better understanding of the underlying dynamics between vertices and facets. Polarity might also serve as a bridge between $\mathcal{H}$-representation (which is easy to build using Proposition IV.3) and $\mathcal{V}$-representation, which will naturally appear for the under-approximations of $\mathcal{R}$ presented in this chapter. The following section introduces such an under-approximation of $\mathcal{R}$ from P, and similarly an over-approximation of $\mathcal{R}^\circ$. Note that the combination of under- and over-approximations, in addition to providing further information about $\mathcal{R}$, can lead to efficient iterative algorithms to approximate $\mathcal{R}$ more precisely, or to determine whether $0 \in \mathcal{R}$.

IV.3 Polytopal under-approximations of a nonempty compact convex set

This section is devoted to the construction of an under-approximation of $\mathcal{R}$ based upon its polytopal over-approximation. We shall first construct it in subsection IV.3.1, then construct an over-approximation of $\mathcal{R}^\circ$ in subsection IV.3.2, ending with conjectures linking the two.

IV.3.1 Under-approximation of $\mathcal{R}$

First, let us define a polytopal under-approximation of $\mathcal{R}$.

Definition IV.6. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set, and a set of directions $(p_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m$. We will consider the under-approximation of $\mathcal{R}$ defined by $(p_j)_j$

and denote $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ the set

$$\mathcal{I}_{\mathcal{R}}((p_j)_j) = \bigcap \{C \subset \mathbb{R}^n \text{ such that } C \text{ is convex and for all } j \in \{1, \dots, m\}, \sigma_C(p_j) = \sigma_{\mathcal{R}}(p_j)\}\,.$$

Thus, $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ denotes the smallest set of the family of convex subsets $C \subset R$ such that $\sigma_C(p_j) = \sigma_{\mathcal{R}}(p_j)$ for every $j \in \{1, \dots, m\}$; these sets share with $\mathcal{R}$ the supporting hyperplanes in the chosen directions, while differing elsewhere.

Proposition IV.7. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set, and a set of directions $(p_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m$, and let P be the sharp over-approximation of $\mathcal{R}$ defined by $(p_j)_j$. Then

$$\mathcal{I}_{\mathcal{R}}((p_j)_j) \subset \mathcal{R} \subset P. \tag{IV.3.1}$$

Proof. Since $\mathcal{R}$ clearly satisfies the requirements of the intersected sets of Definition IV.6, one immediately has $\mathcal{I}_{\mathcal{R}}((p_j)_j) \subset \mathcal{R}$. The second inclusion is exactly the result of Proposition IV.3. □

Remark IV.8. Let us note a few things about $\mathcal{I}_{\mathcal{R}}((p_j)_j)$:

- $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ can be empty (for example if P is a triangle in $\mathbb{R}^2$, then $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is the intersection of all triangles whose vertices are chosen one from each of the three supporting facets, which have an empty intersection) or reduced to a singleton (if P is a rectangle in $\mathbb{R}^2$, then all quadrilaterals having one vertex on each facet of P will contain the centre of P). It follows that $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ will generally not be a sharp polytopal under-approximation of $\mathcal{R}$

- As such, it would just as well be defined if $\mathcal{R}$ is not compact. In that case, $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ would remain bounded since inside the intersection necessarily lies a compact convex set.

For a visualisation in $\mathbb{R}^2$, see Figure IV.2 – note that for visualisation purposes, all figures in this chapter will present approximations of the same example.

Across this section, we will prove that $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is a polytope. First, we will prove the following theorem, which states that $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is the intersection of all polytopes Q included in P such that for each facet F of P, one of the vertices of Q is among the vertices of F.

Theorem IV.9. Let $\mathcal{R}$ be a nonempty compact convex set, and P be a bounded sharp polytopal over-approximation of $\mathcal{R}$. Let $(p_j)_{j\in\{1,\dots,m\}}$ be its $\mathcal{H}$-representation, and $(F_j)_j$ be the associated facets defined by

$$\forall j \in \{1, \dots, m\}, \quad F_j = P \cap \{x \in \mathbb{R}^n, \langle p_j, x\rangle = \sigma_{\mathcal{R}}(p_j)\}. \tag{IV.3.2}$$

Define as well, for each facet F_j, the vertices of P lying in the boundary of F_j by $V_j = (s_j^i)_{i\in\{1,\dots,k_j\}}$. Then

$$\mathcal{I}_{\mathcal{R}}((p_j)_j) = \bigcap_{(s_j)_{j\in\{1,\dots,m\}} \in V_1\times\cdots\times V_m} \operatorname{conv}\{(s_j)_j\}\,. \tag{IV.3.3}$$

It follows that $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is a bounded polytope.

The proof of the theorem relies on the following proposition and lemma: the first proves that $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is the intersection of all polytopes having a vertex on each facet of P.

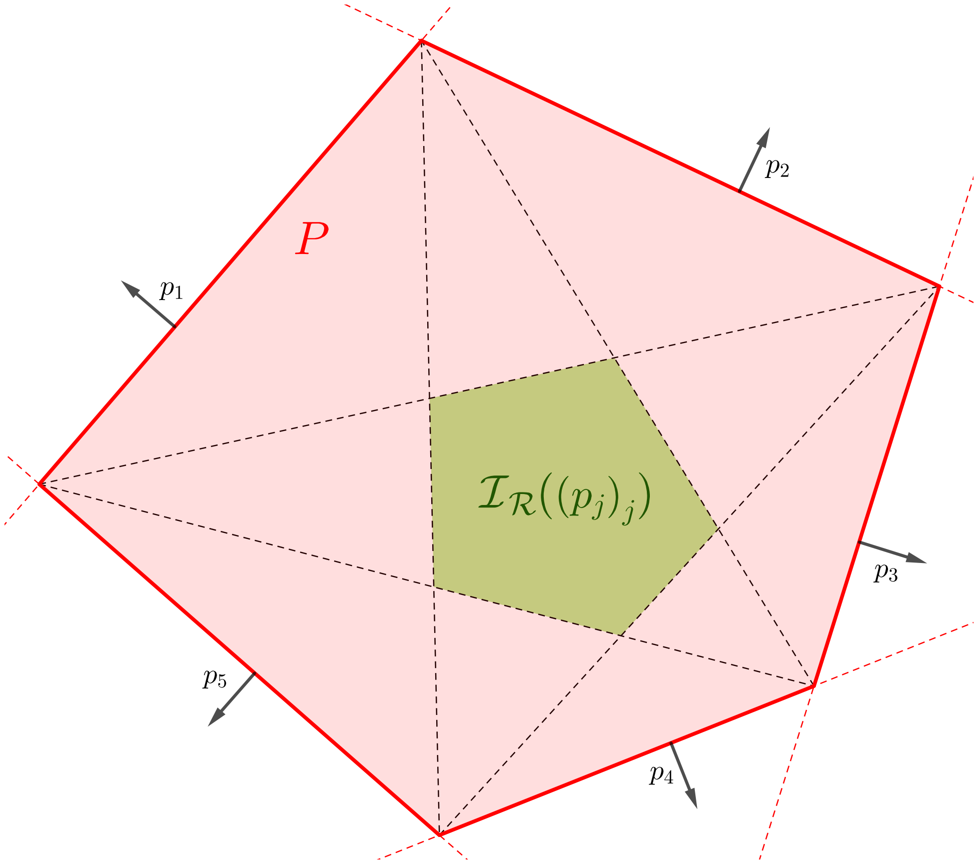


Figure IV.2: Polytopal under-approximation of $\mathcal{R}$.

Proposition IV.10. Let P and $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ be defined as in Definition IV.6, and define facets and vertices as in Theorem IV.9. Then $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is the intersection of polytopes with one vertex in each facet of P, that is,

$$\mathcal{I}_{\mathcal{R}}((p_j)_j) = \bigcap_{(s_j)_{j\in\{1,\dots,m\}}\in F_1\times\cdots\times F_m} \operatorname{conv}\{(s_j)_j\}. \tag{IV.3.4}$$

Proof. For all nonempty compact convex set $C \subset \mathbb{R}^n$, we have the equivalence

$$\forall\, j \in \{1,\dots,m\}, \sigma_C(p_j) = \sigma_{\mathcal{R}}(p_j) \iff C \subset P \text{ and } F_j \cap C \neq \emptyset. \tag{IV.3.5}$$

Consequently, we have that

$$\begin{aligned}
\mathcal{I}_{\mathcal{R}}((p_j)_j) &= \bigcap \{C \subset \mathbb{R}^n,\ \text{s.t. } C \text{ is convex and } \forall\, j \in \{1,\dots,m\}, \sigma_C(p_j) = \sigma_{\mathcal{R}}(p_j)\} \\
&= \bigcap \{C \subset \mathbb{R}^n,\ \text{s.t. } C \text{ is convex and } \forall\, j \in \{1,\dots,m\}, F_j \cap C \neq \emptyset, C \subset P\} \\
&= \bigcap_{(s_j)_j\in F_1\times\cdots\times F_m} P \cap \bigcap_{\substack{(s_j)_j\in C\subset\mathbb{R}^n \\ C \text{ convex}}} C \\
&= \bigcap_{(s_j)_j\in F_1\times\cdots\times F_m} P \cap \operatorname{conv}((s_j)_j) \text{ using Proposition I.32} \\
&= \bigcap_{(s_j)_j\in F_1\times\cdots\times F_m} \operatorname{conv}((s_j)_j).
\end{aligned}$$

□

The second lemma useful to the proof of Theorem IV.9 is the following:

Lemma IV.11. Let A and F be two convex sets such that F is compact. Denote $\mathrm{ext}(F)$ the set of extreme points of F. Then

$$\bigcap_{f\in F} \mathrm{conv}(A\cup f) = \bigcap_{s\in \mathrm{ext}(F)} \mathrm{conv}(A\cup s). \tag{IV.3.6}$$

Proof. Since F is compact, we have that $\mathrm{ext}(F) \subset F$ and $F = \mathrm{conv}(S)$. It follows that

$$\bigcap_{f\in F} \mathrm{conv}(A\cup s) \subset \bigcap_{s\in \mathrm{ext}(F)} \mathrm{conv}(A\cup s). \tag{IV.3.7}$$

Let us prove the converse inclusion. Let $x \in \bigcap_{s\in\mathrm{ext}(F)} \mathrm{conv}(A\cup s)$ and $f \in F$. Proving that $x \in \mathrm{conv}(A\cup f)$ will conclude the demonstration.

$f \in F = \mathrm{conv}(\mathrm{ext}(F))$, therefore using Caratheodory's theorem we know that there exist $(\mu_i)_{i\in\{1,\dots,n+1\}} \in [0,1]^{n+1}$ and $(s_i)_{i\in\{1,\dots,n+1\}} \in \mathrm{ext}(F)^{n+1}$ such that $\sum_{i=1}^{n+1}\mu_i = 1$ and $\sum_{i=1}^{n+1}\mu_i s_i = f$.

Two cases then appear: if $x \in A$, then $x \in \mathrm{conv}(A\cup f)$ and the proof holds. If $x \notin A$, then since for all $s \in \mathrm{ext}(F)$, $x \in \mathrm{conv}(A\cup s)$, we have that for all $i \in \{1,\dots,n+1\}$, there exist $\lambda_i \in [0,1]$ and $a_i \in A$ such that $x = \lambda_i s_i + (1-\lambda_i)a_i$, and furthermore $\lambda_i \neq 0$ because $x \notin A$. Therefore we have that $s_i = \frac{1}{\lambda_i}(x-(1-\lambda_i)a_i)$. It follows that

$$f = \sum_{i=1}^{n+1}\mu_i s_i = \sum_{i=1}^{n+1}\frac{\mu_i}{\lambda_i}(x-(1-\lambda_i)a_i) = x\sum_{i=1}^{n+1}\frac{\mu_i}{\lambda_i} - \sum_{i=1}^{n+1}\frac{\mu_i(1-\lambda_i)}{\lambda_i}a_i, \tag{IV.3.8}$$

and therefore

$$x = \frac{1}{\sum_{i=1}^{n+1}\frac{\mu_i}{\lambda_i}}\left(f + \sum_{i=1}^{n+1}\frac{\mu_i(1-\lambda_i)}{\lambda_i}a_i\right). \tag{IV.3.9}$$

One can then easily check that x is a convex combination of f and the $(a_i)_i$, which means that $x \in \mathrm{conv}(A\cup f)$, concluding the proof. □

Let us now prove Theorem IV.9.

Proof of Theorem IV.9. Using Proposition IV.10, we have that

$$\mathcal{I}_{\mathcal{R}}((p_j)_j) = \bigcap_{(s_j)_{j\in\{1,\dots,m\}}\in F_1\times\cdots\times F_m} \mathrm{conv}\left\{(s_j)_j\right\}. \tag{IV.3.10}$$

Notice that

$$\mathcal{I}_{\mathcal{R}}((p_j)_j) = \bigcap_{(s_j)_{j\in\{2,\dots,m\}}\in F_2\times\cdots\times F_m}\ \bigcap_{s_1\in F_1} \mathrm{conv}\left\{(s_j)_j\right\}. \tag{IV.3.11}$$

Using Lemma IV.11 with $A = \{s_j,\ j\in\{2,\dots,m\}\}$ and $F = F_1 = \mathrm{conv}(V_1)$, we then have that

$$\bigcap_{s_1\in F_1} \mathrm{conv}\left\{A\cup\{s_1\}\right\} = \bigcap_{s_1\in V_1} \mathrm{conv}\left\{A\cup\{s_1\}\right\}, \tag{IV.3.12}$$

and therefore

$$\mathcal{I}_{\mathcal{R}}((p_j)_j) = \bigcap_{(s_j)_{j\in\{1,\dots,m\}}\in V_1\times F_2\times\cdots\times F_m} \mathrm{conv}\left\{(s_j)_j\right\}. \tag{IV.3.13}$$

Reasoning by induction over each facet produces the desired formula. Finally, since all convex hulls of finitely many points are polytopes and that all intersections of finitely many polytopes is a polytope, $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is a polytope. □

This result allows for computation of this under-approximation, for it is now a finite intersection of polytopes. The number of polytopes to intersect is still very large, and increases significantly with the dimension n: when $n = 2$, there is usually 2 (sometimes 1) vertex per facet, so little computation is needed to calculate $\mathcal{I}_{\mathcal{R}}((p_j)_j)$, but already for $n = 3$ it is much more complex, and some more research would have to be done to accelerate its computation.

This intersection would still need to be extensively studied to determine whether it is interesting and useful. For instance, we believe that, like its polar set, it has one vertex for each direction p_j – or at least for every $(n-1)$-dimensional facet of P. It may also have links to the intersection of the polytopes having all the same vertices as the over-approximations except one – this can clearly be seen from the two-dimensional example in Figure IV.2, but it fails when a facet is reduced to a singleton, or in higher dimensions. All these properties would be crucial to the development of iterative methods of approximation, which we discuss in Section IV.5.

One lead to tackle these problems would be to use the polar representation of $\mathcal{I}_{\mathcal{R}}((p_j)_j)$, which we try to characterise in the following subsection.

IV.3.2 Over-approximation of $\mathcal{R}^\circ$

Recall that for $(p_j)_j$ and $(\sigma_{\mathcal{R}}(p_j))_j$ that define a polytopal over-approximation P of $\mathcal{R}$, we have that P° is a polytopal under-approximation of $\mathcal{R}^\circ$. Similarly to under-approximations of $\mathcal{R}$, one can wonder whether it is possible to compute over-approximations of $\mathcal{R}^\circ$ using only the data $(p_j)_j$ and $(\sigma_{\mathcal{R}}(p_j))_j$. This is the topic of this subsection.

Definition IV.12. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set. Let $(p_j)_{j\in\{1,\dots,m\}}$. We will consider the over-approximation of $\mathcal{R}^\circ$ defined by $(p_j)_j$ and denote $\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)$ the set

$$\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j) = \bigcup \left\{ C \subset \mathbb{R}^n \text{ is convex}, P^\circ \subset C, \exists\, (q_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m, \sigma_C(q_j) = \langle q_j, \tfrac{p_j}{\sigma_{\mathcal{R}}(p_j)} \rangle \right\},$$

that is, the union of all convex sets C having each vertex of P° on its boundary.

Notice that this union, although clearly including $\mathcal{R}$, is not convex, and may not be bounded: this is the case, for instance, if in $\mathbb{R}^2$ four or fewer directions $(p_j)_j$ are considered. An illustration of $\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)$ can be seen in Figure IV.3.

Similarly to Proposition IV.10, one can only consider polytopes in the union defining $\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)$:

Proposition IV.13. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set. Let $(p_j)_{j\in\{1,\dots,m\}}$. For $p, q \in \mathbb{R}^n$, denoting the half-space

$$H_q(p) = \{x \in \mathbb{R}^n, \langle q, x\rangle \le \langle q, p\rangle\}, \tag{IV.3.14}$$

we have that

$$\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j) = \bigcup_{(q_j)_{j\in\{1,\dots,m\}}\in(\mathbb{R}^n)^m} \ \bigcap_{j\in\{1,\dots,m\}} H_{q_j}\left(\frac{p_j}{\sigma_{\mathcal{R}}(p_j)}\right), \tag{IV.3.15}$$

that is, $\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)$ is the union of all polytopes containing P° and having $(\frac{p_j}{\sigma_{\mathcal{R}}(p_j)})_j$ at their boundary.

Proof. Given $(q_j)_{j\in\{1,\dots,m\}}$ and $C \subset \mathbb{R}^n$, we have that

$$\left[\forall\, j \in \{1,\dots,m\}, \sigma_C(q_j) = \langle q_j, \tfrac{p_j}{\sigma_{\mathcal{R}}(p_j)}\rangle\right] \iff C \subset \bigcap_{j\in\{1,\dots,m\}} H_{q_j}\left(\frac{p_j}{\sigma_{\mathcal{R}}(p_j)}\right). \tag{IV.3.16}$$

Indeed, the right-hand side set contains all sets satisfying the left-hand side's conditions, while also satisfying them. It follows that it is the union of all such sets C (and in particular of P°)

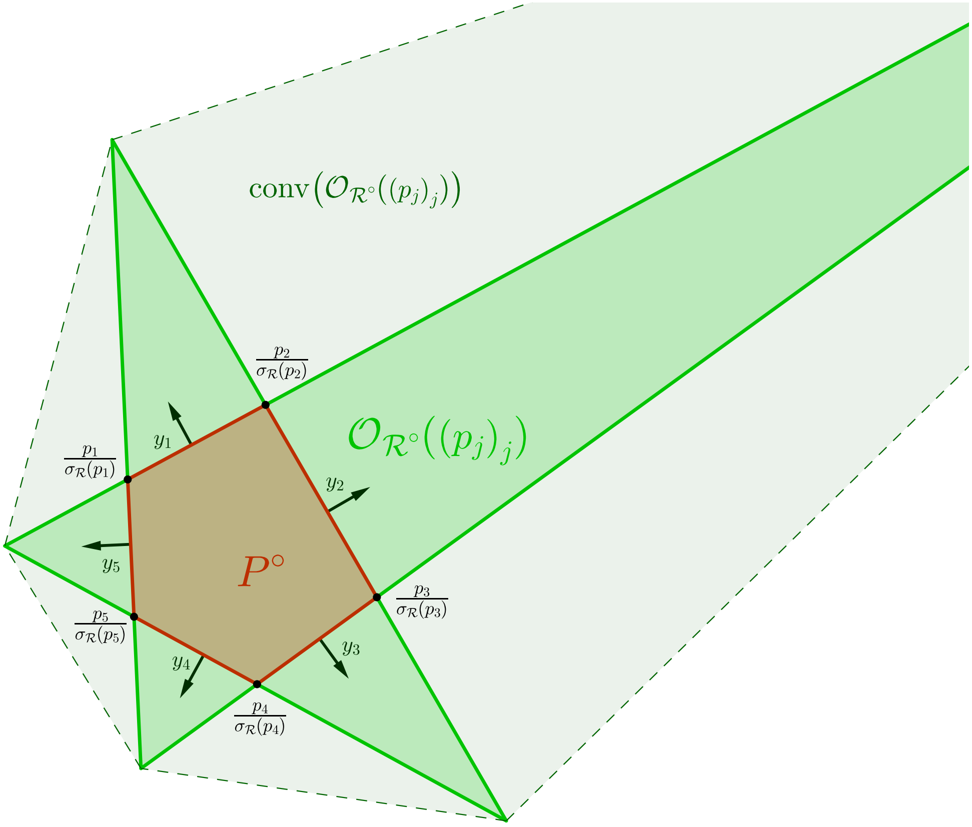

$\mathrm{conv}\big(\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)\big)$
$\frac{p_2}{\sigma_{\mathcal{R}}(p_2)}$
y_1
$\frac{p_1}{\sigma_{\mathcal{R}}(p_1)}$
$\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)$
y_2
P°
y_5
$\frac{p_3}{\sigma_{\mathcal{R}}(p_3)}$
$\frac{p_5}{\sigma_{\mathcal{R}}(p_5)}$
y_4
y_3
$\frac{p_4}{\sigma_{\mathcal{R}}(p_4)}$

Figure IV.3: Over-approximation of $\mathcal{R}^\circ$.

and thus that

$$\begin{aligned}\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j) &= \bigcup_{(q_j)_{j\in\{1,\dots,m\}}\in(\mathbb{R}^n)^m} \left\{C \subset \mathbb{R}^n \text{ is convex}, P^\circ \subset C, \sigma_C(q_j) = \langle q_j, \tfrac{p_j}{\sigma_{\mathcal{R}}(p_j)}\rangle\right\} \\ &= \bigcup_{(q_j)_{j\in\{1,\dots,m\}}\in(\mathbb{R}^n)^m} \bigcap_{j\in\{1,\dots,m\}} H_{q_j}\left(\tfrac{p_j}{\sigma_{\mathcal{R}}(p_j)}\right).\end{aligned}$$

□

This expression, though more tractable than the definition of $\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)$, still incorporates a union over an infinite – uncountable, even – number of directions. We believe that, similarly to the expression of $\mathcal{I}_{\mathcal{R}}(p_j)_j$ developed in Theorem IV.9, one could simplify it to a finite number of directions – for each vertex, the directions of its adjacent facets. This can clearly be seen in the illustration in $\mathbb{R}^2$ presented in Figure IV.3.

Going even further, we expect to be able to achieve the following conjecture, proving the duality between the under-approximation of $\mathcal{R}$ and the over-approximation of $\mathcal{R}^\circ$.

Conjecture IV.14. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set, and let $(p_j)_{j\in\{1,\dots,m\}}$. We then have

$$\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)^\circ = \mathrm{conv}(\{0\} \cup \mathcal{I}_{\mathcal{R}}((p_j)_j)) \qquad \text{and} \qquad \mathcal{I}_{\mathcal{R}}((p_j)_j)^\circ = \mathrm{conv}\left(\mathcal{O}_{\mathcal{R}^\circ}((p_j)_j)\right). \quad \text{(IV.3.17)}$$

Another conjecture we have would concern an inverse-like property between over- and under-approximations of this form: in Figure IV.3, it can clearly be seen that P° could be built from $\mathcal{O}_{\mathcal{R}}((p_j)_j)$ similarly to the way $\mathcal{I}_{\mathcal{R}}((p_j)_j)$ is constructed from P in Figure IV.2.

Conjecture IV.15. Let P a bounded polytope. Let us denote somewhat abusively the under-approximations of $\mathcal{R}$ and over-approximations of $\mathcal{R}^\circ$ as under- and over-approximations of the polytopes: $\mathcal{I}(Q)$ and $\mathcal{O}(Q)$ for all polytopes Q. We then have an inverse-like property between under- and over-approximations:

$$\mathcal{O}(\mathcal{I}(P)) = P. \qquad \text{(IV.3.18)}$$

Furthermore, if $\mathcal{O}(P)$ is bounded, we have that

$$\mathcal{I}(\mathcal{O}(P)) = P. \tag{IV.3.19}$$

Such results would allow for a better understanding of the under-approximation of $\mathcal{R}$, which could then lead to interesting ways to compute approximations of $\mathcal{R}$, or to determine if $0 \in \mathcal{R}$. In the following section, we consider the case where we have access to further information on $\mathcal{R}$.

IV.4 Under-approximation of $\mathcal{R}$ with maximisers of $\sigma_{\mathcal{R}}$

In the previous section, we mainly considered approximations of a nonempty compact convex set $\mathcal{R} \subset \mathbb{R}^n$ when having access only to evaluations of the support function at given directions $(p_j)_j$. In this section, we shall assume to have a little bit more information: recall that the support function of $\mathcal{R}$ reads

$$\forall\, p \in \mathbb{R}^n, \quad \sigma_{\mathcal{R}}(p) = \sup_{r \in \mathcal{R}} \langle p, r \rangle. \tag{IV.4.1}$$

Since $\mathcal{R}$ is assumed to be compact and the inner product is continuous with respect to its input, this supremum is reached and therefore

$$\forall\, p \in \mathbb{R}^n,\ \exists\, r \in \mathcal{R}, \quad \sigma_{\mathcal{R}}(p) = \langle p, r \rangle. \tag{IV.4.2}$$

In most configurations where a support function is computable, it is so using the knowledge of such a maximiser $r \in \mathcal{R}$ – indeed, in all the examples considered in this thesis, this was the case. Of course, it is always known up to some error (discretisation, rounding, etc.), but let us assume within this section that we have access to such maximisers, leaving the error-case for Section IV.6.

Proposition IV.16. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set. Let $(p_j)_{j \in \{1,\ldots,m\}}$ be a set of directions, and let $(r_j)_{j \in \{1,\ldots,m\}} \in \mathcal{R}^m$ be such that

$$\forall\, j \in \{1, \ldots, m\}, \quad \sigma_{\mathcal{R}}(p_j) = \langle p_j, r_j \rangle. \tag{IV.4.3}$$

Then, denoting P the sharp polytopal approximation of $\mathcal{R}$ defined by the directions $(p_j)_j$ we have the inclusions

$$\operatorname{conv}((r_j)_j) \subset \mathcal{R} \subset P. \tag{IV.4.4}$$

Since this proposition's proof only relies on the convexity of $\mathcal{R}$ and on Proposition IV.3, we will not detail it. The under-approximation of $\mathcal{R}$ is larger than $\mathcal{I}((p_j)_j)$: indeed, recall that

$$\mathcal{I}((p_j)) = \bigcap \{C \subset \mathbb{R}^n, C \text{ convex s.t. } \forall\, j \in \{1, \ldots, m\}, \sigma_C(p_j) = \sigma_{\mathcal{R}}(p_j)\}. \tag{IV.4.5}$$

In particular, $\operatorname{conv}((r_j)_j)$ is one of such convex sets, which means that $\mathcal{I}((p_j)_j) \subset \operatorname{conv}((r_j)_j)$. In fact, recalling Definition IV.4, $\operatorname{conv}((r_j)_j)$ defined by Proposition IV.16 is a sharp under-approximation of $\mathcal{R}$: an illustration of $\operatorname{conv}((r_j)_j)$ and P is provided in Figure IV.4.

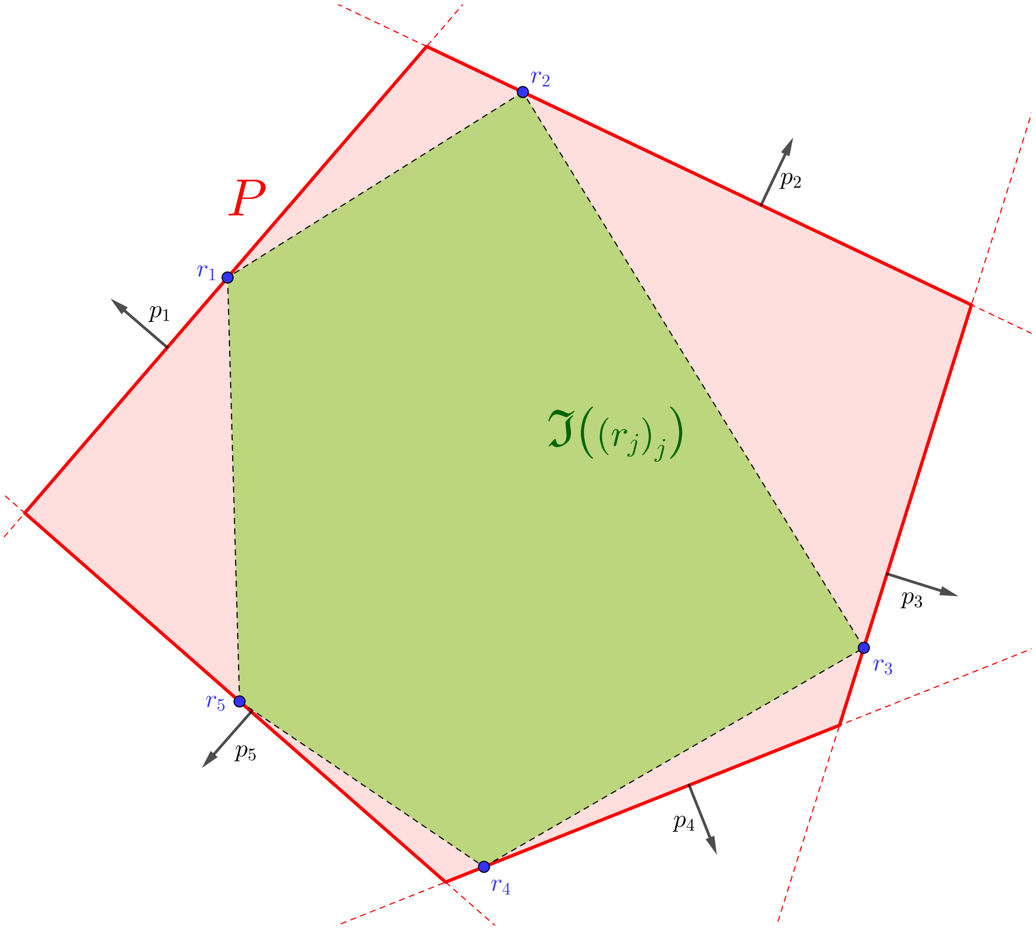


Figure IV.4: Polytope under-approximation of $\mathcal{R}$ with maximisers.

One can also wonder what this would correspond to as approximations of $\mathcal{R}^\circ$. The following proposition proves that $\operatorname{conv}((r_j)_j)$'s polar is a sharp polytopal over-approximation of $\mathcal{R}^\circ$

Proposition IV.17. Let $\mathcal{R} \subset \mathbb{R}^n$ a nonempty compact convex set. Let $(p_j)_{j\in\{1,\dots,m\}} \in (\mathbb{R}^n)^m$ define a sharp over-approximation P of $\mathcal{R}$ such that $0 \in \operatorname{Int}(P)$ and let $(r_j)_{j\in\{1,\dots,m\}} \in (\mathcal{R})^m$ such that

$$\forall\, j \in \{1,\dots,m\}, \quad \sigma_{\mathcal{R}}(p_j) = \langle p_j, r_j\rangle. \tag{IV.4.6}$$

Denoting

$$\mathfrak{I}((r_j)_j) = \operatorname{conv}((r_j)_j), \tag{IV.4.7}$$

we have that $\mathfrak{I}((r_j)_j)^\circ$ is a sharp polytopal over-approximation of $\mathcal{R}^\circ$.

Proof. Proposition I.43 yields that $\mathfrak{I}((r_j)_j)^\circ$ is a polytope defined by its $\mathcal{H}$-representation $(r_j)_j$ and $(b_j)_j = (1)_j$. Since $\mathfrak{I}((r_j)_{j\in\{1,\dots,m\}}) \subset \mathcal{R}$, we have that $\mathcal{R}^\circ \subset \mathfrak{I}((r_j)_{j\in\{1,\dots,m\}})^\circ$. Therefore, the only thing left to prove is that for all $j \in \{1,\dots,m\}$, $\sigma_{\mathcal{R}^\circ}(r_j) = 1$.

Let $j \in \{1,\dots,m\}$. By Theorem I.35, $r_j \in \mathcal{R} \subset \mathcal{R}^{\circ\circ} = (\mathcal{R}^\circ)^\circ$ and thus $\sigma_{\mathcal{R}^\circ}(r_j) \le 1$. Furthermore, since $0 \in \operatorname{Int} P$, $\sigma_{\mathcal{R}}(p_j) = \sigma_P(p_j) > 0$, we have that

$$\sigma_{\mathcal{R}}\left(\frac{p_j}{\sigma_{\mathcal{R}}(p_j)}\right) = 1 \tag{IV.4.8}$$

and therefore $\frac{p_j}{\sigma_{\mathcal{R}}(p_j)} \in \mathcal{R}^\circ$. This yields

$$\sigma_{\mathcal{R}^\circ}(r_j) = \sup_{x\in\mathcal{R}^\circ} \langle r_j, x\rangle \ge \left\langle r_j, \frac{p_j}{\sigma_{\mathcal{R}}(p_j)}\right\rangle = 1, \tag{IV.4.9}$$

which concludes the proof that $\sigma_{\mathcal{R}^\circ}(r_j) = 1$ and that $\mathfrak{I}((r_j)_{j\in\{1,\ldots,m\}})^\circ$ is a sharp polytopal over-approximation of $\mathcal{R}^\circ$. □

Figure IV.5 presents an illustration of such an over-approximation of $\mathcal{R}^\circ$.

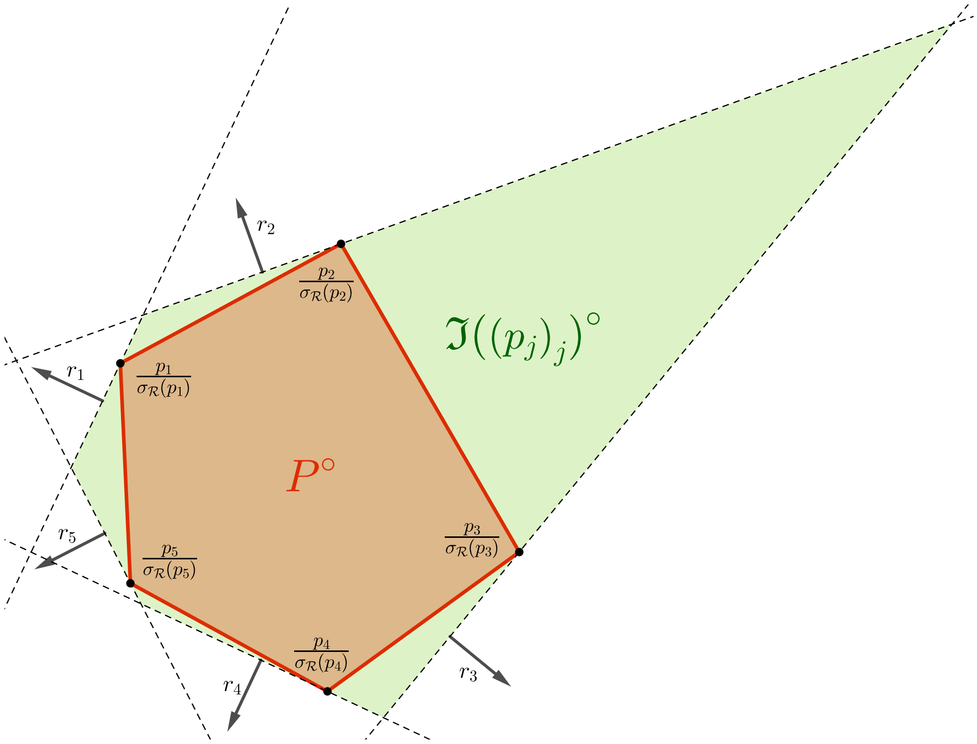

r_2
$\frac{p_2}{\sigma_{\mathcal{R}}(p_2)}$
$\mathfrak{I}((p_j)_j)^\circ$
r_1
$\frac{p_1}{\sigma_{\mathcal{R}}(p_1)}$
P°
$\frac{p_3}{\sigma_{\mathcal{R}}(p_3)}$
r_5
$\frac{p_5}{\sigma_{\mathcal{R}}(p_5)}$
$\frac{p_4}{\sigma_{\mathcal{R}}(p_4)}$
r_4
r_3

Figure IV.5: Over-approximation of $\mathcal{R}^\circ$ with maximisers.

In the following section, we consider iterative methods to efficiently compute under- and over-approximations, as well as deciding whether $0 \in \mathcal{R}$.

IV.5 Iterative methods

Recall that we had originally two objectives when evaluating the support function of $\mathcal{R}$:

- a first goal was to decide whether $0 \in \mathcal{R}$ – equivalent to the reachability of a single target;
- a more complex one was to approximate $\mathcal{R}$ – equivalent to approximating the reachable set.

Across the previous sections, we have focused on the second objective, first approximating the reachable set from outside in Section IV.2, then from inside in the following ones. We shall therefore first discuss methods of iteratively producing under- and over-approximations of $\mathcal{R}$, then see how we could adapt these methods to prove that $0 \in \mathcal{R}$.

Before delving into a few ideas of algorithms, it is important to reassert that this section (along with others in this chapter) does not present finished work nor claim authenticity of its results: approximations of convex bodies, and in particular using inner and outer polytopes, have been extensively studied. We refer the reader to the literature for a more academic and rigorous view of the matter, for example in the survey [Bro08].

First, one can notice that to produce an under-approximation $\mathcal{I}$ of $\mathcal{R}$, and by extension to prove that $0 \in \mathcal{R}$, one has to provide global information on its support function and therefore the number of directions $(p_j)_{j\in\{1,\ldots,m\}}$ considered must be high enough. When one has access to the maximisers of the support function, immediately those maximisers $(r_j)_{j\in\{1,\ldots,m\}}$ are proved to be in $\mathcal{R}$. However, one can only prove that a set $\operatorname{conv}((r_j)_{j\in\{1,\ldots,m\}})$ of nonzero measure if $m \geq n+1$.

Similarly, we believe that if the maximisers are not known, a necessary condition to obtain a nonempty under-approximation is that $m \geq 2n$. For instance, if $(e_i)_i$ is the canonical basis of $\mathbb{R}^n$, then evaluating $\sigma_{\mathcal{R}}$ on each e_i and $-e_i$ will provide a hyper-rectangle over-approximation and prove that its centre (and its centre only) is included in $\mathcal{R}$. Surprisingly enough, if the over-approximation produced by the $(p_j)_j$ is unbounded, even with $2n$ directions one can prove the inclusion of a non-singleton set $\mathcal{I}$ (although still of measure zero).

Once an under-approximation has been constructed, one can compare under- and over-approximation to choose new directions to pursue more precise approximations. One could also choose all the directions $(p_j)_j$ in advance, and then compute both approximations, but doing so iteratively has multiple advantages:

- of course, it allows one to adapt the strategy at each iteration: if the computation of each $\sigma_{\mathcal{R}}(p_j)$ (and potentially the associated maximiser) is expensive, this makes even more sense – and this is the case in this thesis, for the computer-assisted proofs require interval arithmetic and very fine discretisation. In particular, if the only purpose is to prove that $0 \in \mathcal{R}$, an iterative method might be very well suited.

- as was seen in the previous sections, computing under-approximations might require the computation of the vertices of the over-approximation. This is a very costly enterprise, which requires knowing which facets border each other. While the prior information and choice of the $(p_j)_j$ will considerably help, computing the vertices iteratively will also alleviate the computational cost.

When considering iterative methods to build an under- and over-approximation of $\mathcal{R}$ based on the inner polytope: since in both cases (with or without maximisers of the support function), we believe that there is one vertex for each facet of the outer polytope, one could consider choosing a normal direction to one of the inner polytope's facets. The choice of the facet could be done using various criteria, such as area of the facet (a large facet has greater potential for a significant improvement) or the distance between the inner and outer polytope. Since that topic has been extensively studied in the literature (see [Bro08, Section 8]), we shall focus hereafter on iterative methods of proving that $0 \in \mathcal{R}$.

If however the purpose is solely to prove or disprove $0 \in \mathcal{R}$, different methods could be envisioned: the naive approach would be of course to simply try and under- and over-approximate $\mathcal{R}$ as well as possible, hoping to obtain $0 \in \mathcal{I}$ or $0 \notin P$. A more subtle one could be to first build a crude under- and over-approximation using a given set of directions, and then build up the pair using an iterative method, for instance choosing the next direction as the opposite of the orthogonal projection of 0 on $\mathcal{I}$, progressively closing in on 0 until inclusion in $\mathcal{I}$ or P^C is proved.

A more sophisticated method would be to first choose all the directions using a descent algorithm on the support function – for instance, a primal-dual algorithm such as the Chambolle-Pock algorithm [CP11] would provide both directions and compute the maximisers at each iteration. Such a descent algorithm would either find a negative point of the support function (and thus imply that $0 \notin \mathcal{R}$), and building the over- and under-approximation at each step would prove, if it is the case, that $0 \in \mathcal{I}$. Although seemingly more efficient, this could also prove more computationally heavy – the algorithm can take many small steps with little impact on the approximations, which are expensive to compute – or even not provide the desired result if in fact $0 \in \mathcal{R}$. For instance, the algorithm could get stuck in a direction leading to the minimiser of $\sigma_{\mathcal{R}}$ (0, in that case), and not provide enough global information to prove its inclusion. To prevent that scenario, one could then switch back from this algorithm to another one, better suited to the $0 \in \mathcal{R}$ case.

Finally, another method could rely on polar sets: recall that for a compact convex set C, $0 \in \mathrm{Int}(C)$ if and only if its polar C° is bounded. It goes the same for $\mathcal{R}$ or its under-approximations $\mathcal{I}$. Therefore, one could search for directions for which the polar of the inner approximation (or the over-approximation $\mathcal{O}$ of $\mathcal{R}^\circ$, see Conjecture IV.14) is not bounded. This setting is, of course, equivalent to the original under-approximation of $\mathcal{R}$, but can still provide useful insights or computation techniques using both primal and dual viewpoints.

IV.6 Polytopal approximations with support function errors

In light of applications to computer-assisted proofs of reachability, one can wonder what impact inaccurate support functions evaluations would have. We shall in this section give an overview of how those errors could affect the computation of the various polytopal approximations surveyed in this chapter.

Across this chapter, we have made the assumption of having an exact evaluation method for the support function, that is, being able to compute $\sigma_{\mathcal{R}}(p)$ for all $p \in \mathbb{R}^n$. In practice, as has been emphasised on multiple occasions throughout the thesis, we only have access to an approximation of it, with explicit bounds on the error. Therefore, in this section we shall only assume approximate knowledge of the support function of $\mathcal{R}$, that is, for all $p \in \mathbb{R}^n$, the ability to compute $b \in \mathbb{R}$ and $e > 0$ such that

$$|\sigma_{\mathcal{R}}(p) - b| \leq e, \tag{IV.6.1}$$

where $\mathcal{R} \subset \mathbb{R}^n$ is, as always, a nonempty compact convex set. Notice that since the directions of $(p_j)_j$ used to construct the polytope have been chosen, the only errors lie in the value of the support function, and, if it is known, in the computation of its maximiser.

All throughout this section, we will use an overline $\overline{x}$ (or $\overline{X}$ for a set) to indicate that a value is guaranteed despite the presence of errors: for example, if P denotes an over-approximation of $\mathcal{R}$, then $\overline{P} \supset P$ will denote a computation of P with errors – necessarily, it will be less precise.

We will now provide details as to how this affects approximations, first focusing on over-approximations and then considering under-approximations.

IV.6.1 Over-approximation of $\mathcal{R}$ and under-approximation of $\mathcal{R}^\circ$

As can be seen using the following proposition, computing over-approximations of $\mathcal{R}$ (or equivalently under-approximations of $\mathcal{R}^\circ$) is not severely impeded by support function errors.

Proposition IV.18. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set, and let $(p_j)_{j\in\{1,\ldots,m\}} \in (\mathbb{R}^n)^m$. Assuming we have computed the $(\sigma_{\mathcal{R}}(p_j))_j$ up to a given error term $(e_j)_j$. We thus have access to

$$\forall j \in \{1, \ldots, m\}, \sigma_{\mathcal{R}}(p_j) \in [b_j - e_j, b_j + e_j]. \tag{IV.6.2}$$

Then the polytope $\overline{P}$ defined by

$$\overline{P} = \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n, \langle p_j, x\rangle \leq b_j + e_j\} \tag{IV.6.3}$$

is a polytopal over-approximation of $\mathcal{R}$.

Therefore, one can also compute over-approximations of compact reachable sets if the support function is known up to an error – see Figure IV.6 for an illustration of the method. These over-approximations are, however, not sharp: in general, $\mathcal{R}$ does not touch the facets of $\overline{P}$, but is within a distance of $\frac{e_j}{\|p_j\|}$ of them.

Similarly, one can compute under-approximations of $\mathcal{R}^\circ$ with support function errors.

Proposition IV.19. Using the same notations as in Proposition IV.18, we have that

$$\overline{P^\circ} = \operatorname{conv}\left(\left(\frac{p_j}{b_j + e_j}\right)_{j\in\{1,\ldots,m\}}\right) \subset \mathcal{R}^\circ. \tag{IV.6.4}$$

An illustration of such an under-approximation can be seen in Figure IV.7.

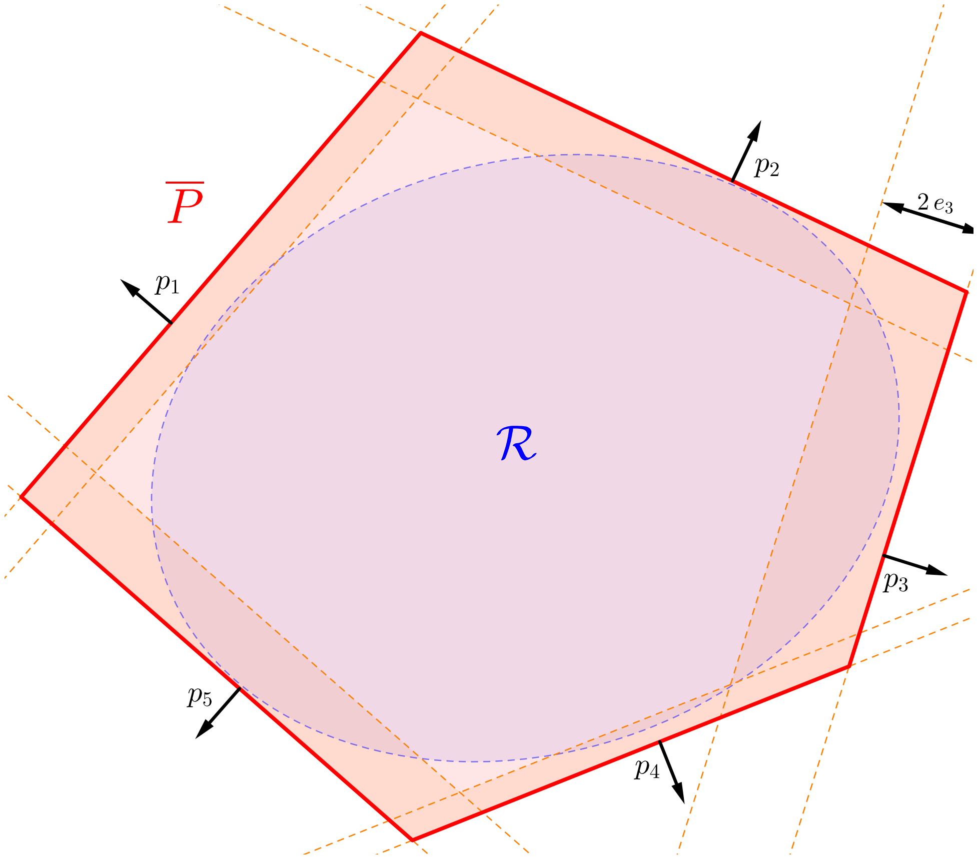


Figure IV.6: Outer polytope approximation with support function errors.

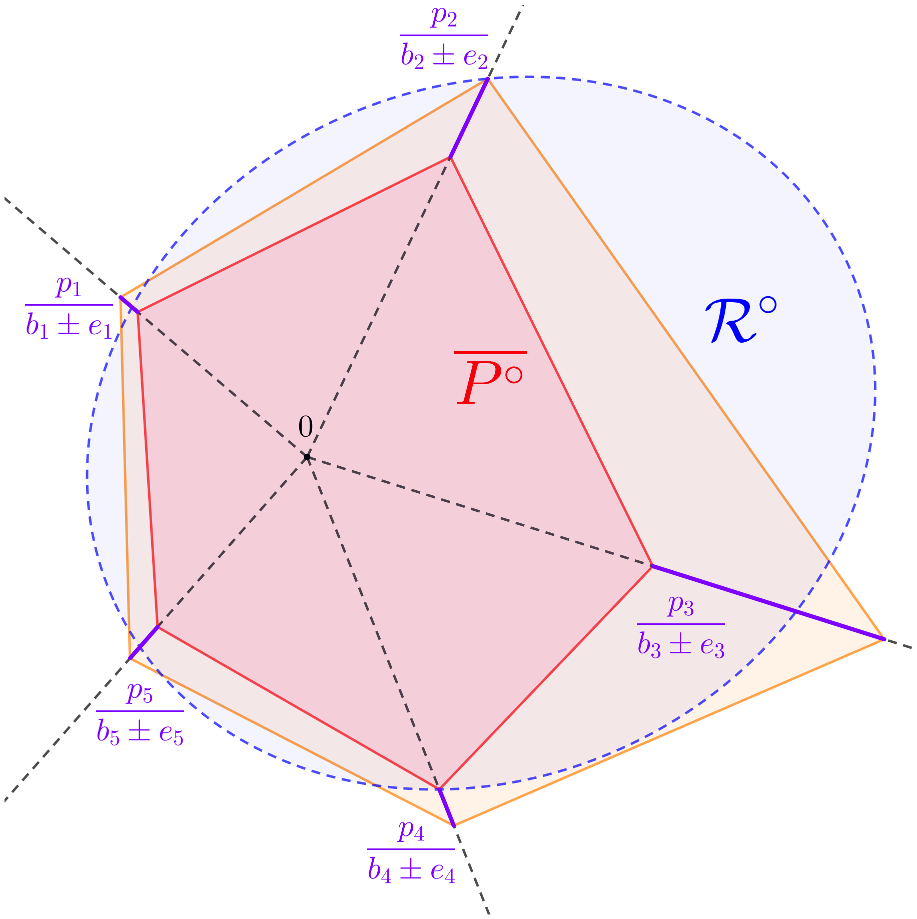

$\frac{p_2}{b_2 \pm e_2}$
$\frac{p_1}{b_1 \pm e_1}$
$\overline{P}^\circ$
$\mathcal{R}^\circ$
0
$\frac{p_3}{b_3 \pm e_3}$
$\frac{p_5}{b_5 \pm e_5}$
$\frac{p_4}{b_4 \pm e_4}$

Figure IV.7: Polytopal under-approximation of $\mathcal{R}^\circ$ with support function errors.

IV.6.2 Under-approximation of $\mathcal{R}$

Computing under-approximations of $\mathcal{R}$ when the computation of $\sigma_\mathcal{R}$ is subject to errors is much more difficult. Since over-approximating $\mathcal{R}^\circ$ poses similar problems, we shall focus here on under-approximations of $\mathcal{R}$.

Assuming first no prior knowledge of the maximisers, recall that Theorem IV.9 characterises the under-approximation $\mathcal{I}_\mathcal{R}((p_j)_j)$ as a finite intersection of convex hulls of the vertices of the over-approximation P. This is no longer the case when considering errors on the support function: in Figure IV.8, we have plotted the new under-approximation $\overline{\mathcal{I}_\mathcal{R}((p_j)_j)}$, computed using a "worst-case" method: the vertices of the sharp over-approximation based on the $(p_j)_j$ lie somewhere within each orange-dotted lozenge on each corner of the pentagon, and therefore when computing the $\overline{\mathcal{I}_\mathcal{R}((p_j)_j)}$, one has to consider all those new possibilities of vertices. However, this illustration is of a simple case: in practice, if more directions are considered, if the errors are higher or if the space dimension is higher, one might witness the appearance of new vertices, the disappearance of others. Therefore, we shall only formulate the following statement about the construction of an under-approximation of $\mathcal{R}$.

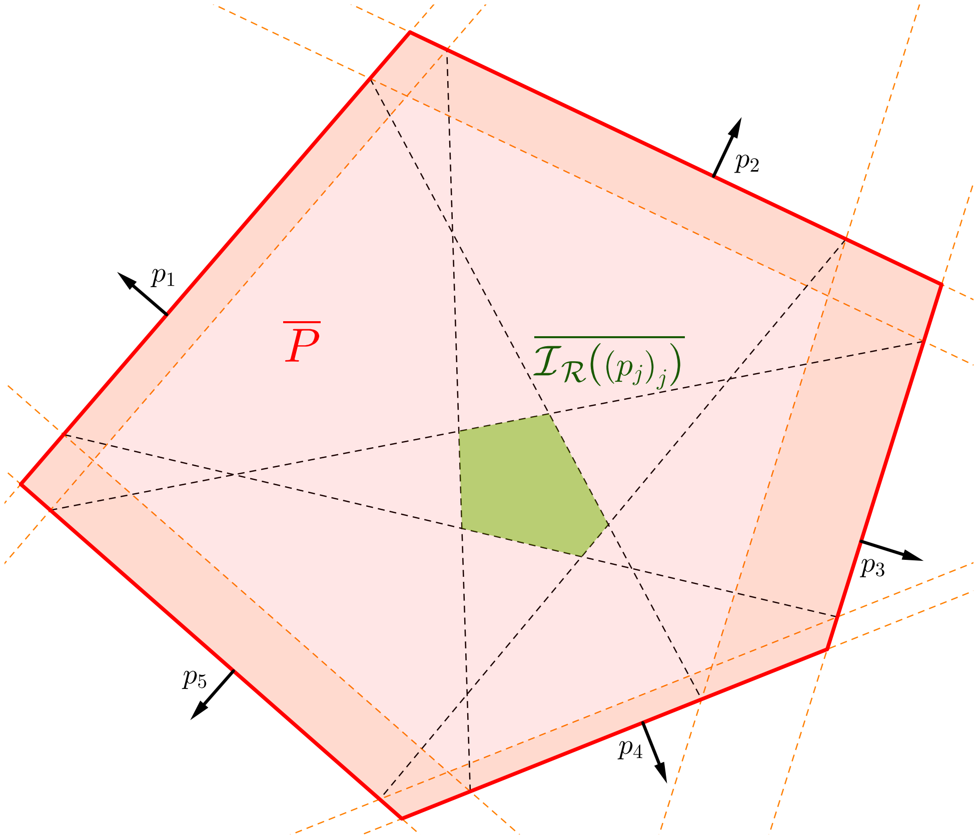


Figure IV.8: Polytopal under- and over-approximations with support function errors.

Proposition IV.20. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set, and let $(p_j)_j$ a set of directions in $\mathbb{R}^n$. Assume that we have computed $(b_j)_j$ and $(e_j)_j$ such that for all j, $|b_j - \sigma_\mathcal{R}(p_j)| \leq e_j$. Denoting

$$\overline{\mathcal{I}_\mathcal{R}((p_j, b_j, e_j)_j)} = \bigcap \{C \text{ convex}, \forall j,\ |\sigma_C(p_j) - b_j| \leq e_j\}\,. \tag{IV.6.5}$$

Then $\overline{\mathcal{I}_\mathcal{R}((p_j, b_j, e_j)_j)} \subset \mathcal{R}$.

The proof of this proposition is immediate, since $\mathcal{R}$ is among the convex sets in the intersection. In the spirit of Proposition IV.16, we believe we could easily restrict the intersection to polytopes. Nevertheless, it is clear that such an intersection will be extremely complicated to compute in practice. Remark that counterintuitively, $\overline{\mathcal{I}_\mathcal{R}((p_j, b_j, e_j)_j)}$ cannot be constructed directly from the "worst-case in each direction", that is, from the polytope

$$\underline{P} = \bigcap_{j=1}^{m} \{x, \quad \langle p_j, x \rangle \leq b_j - e_j\} . \tag{IV.6.6}$$

This can clearly be seen in Figure IV.8.

Were one in possession of the maximisers of the support functions, albeit also computed with errors, one could still provide decent under-approximations of $\mathcal{R}$. Figure IV.9 presents such a case.

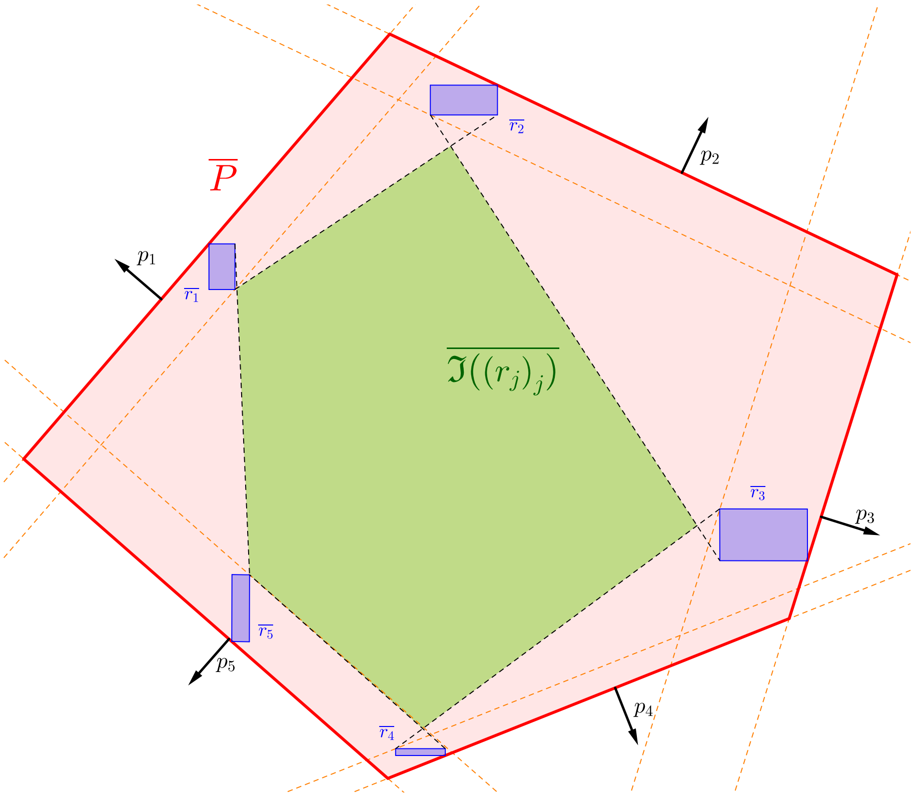


Figure IV.9: Polytopal under- and over-approximations with maximisers and support function errors.

One easily notices that once again an intersection over potential candidates is required to compute an under-approximation. This is the topic of the following proposition, whose proof is immediate.

Proposition IV.21. Let $\mathcal{R} \subset \mathbb{R}^n$ be a nonempty compact convex set, and for all $j \in \{1, \ldots, m\}$, let $\overline{r_j} \subset \mathbb{R}^n$ such that

$$\forall j \in \{1, \ldots, m\}, \quad \overline{r_j} \cap \mathcal{R} \neq \emptyset. \tag{IV.6.7}$$

Then

$$\mathfrak{I}((\overline{r_j})_j) = \bigcap_{(r_j)_j \in \overline{r_1} \times \cdots \times \overline{r_m}} \mathrm{conv}((r_j)_j) \tag{IV.6.8}$$

satisfies $\mathfrak{I}((\overline{r_j})_j) \subset \mathcal{R}$.

In particular, if those $(\overline{r_j})_j$ are constructed as the maximisers of the support function for a set of directions $(p_j)_j$, one is assured that $\overline{r_j} \cap \partial\mathcal{R}$ is nonempty – which is the case pictured in Figure IV.9. If those $(\overline{r_j})_j$ are furthermore convex, we believe that, similarly to Theorem IV.9, one could restrict the intersection to the extreme points of each $\overline{r_j}$, allowing for easier computation of the under-approximation of $\mathcal{R}$.

Bibliography

[Ahm85] NU Ahmed. “Finite-time null controllability for a class of linear evolution equations on a Banach space with control constraints”. In: *Journal of optimization Theory and Applications* 47.2 (1985), pp. 129–158 (cit. on p. 10).

[AFG21] Matthias Althoff, Goran Frehse, and Antoine Girard. “Set Propagation Techniques for Reachability Analysis”. In: *Annual Review of Control, Robotics, and Autonomous Systems* 4 (2021). Publisher: Annual Reviews, pp. 369–395 (cit. on pp. 8, 66, 87).

[ASB10] Matthias Althoff, Olaf Stursberg, and Martin Buss. “Computing reachable sets of hybrid systems using a combination of zonotopes and polytopes”. In: *Nonlinear analysis: hybrid systems* 4.2 (2010), pp. 233–249 (cit. on p. 8).

[Ame+16] Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. “Control barrier function based quadratic programs for safety critical systems”. In: *IEEE Transactions on Automatic Control* 62.8 (2016), pp. 3861–3876 (cit. on p. 8).

[Ant+24] Harbir Antil, Umberto Biccari, Rodrigo Ponce, Mahamadi Warma, and Sebastián Zamorano. “Controllability properties from the exterior under positivity constraints for a 1-D fractional heat equation”. In: *Evolution Equations and Control Theory* 13.3 (2024), pp. 893–924 (cit. on pp. 9, 87).

[Bai+07] R. Baier, C. Büskens, I. A. Chahma, and M. Gerdts. “Approximation of reachable sets by direct solution methods for optimal control problems”. In: *Optimization Methods and Software* 22.3 (2007). Publisher: Taylor & Francis, pp. 433–452 (cit. on pp. 8, 66).

[Bal+18] István Balázs, Jan Bouwe van den Berg, Julien Courtois, János Dudás, Jean-Philippe Lessard, Anett Vörös-Kiss, JF Williams, and Xi Yuan Yin. “Computer-assisted proofs for radially symmetric solutions of PDEs”. In: *Journal of Computational Dynamics* 5.1&2 (2018), pp. 61–80 (cit. on pp. 10, 87).

[BLR92] Claude Bardos, Gilles Lebeau, and Jeffrey Rauch. “Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary”. In: *SIAM journal on control and optimization* 30.5 (1992), pp. 1024–1065 (cit. on p. 9).

[BC11] Heinz Bauschke and Patrick Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Space.* Jan. 2011 (cit. on pp. 35, 38, 42).

[Bea+18] Logan D. R. Beal, Daniel C. Hill, R. Abraham Martin, and John D. Hedengren. “GEKKO Optimization Suite”. In: *Processes* 6.8 (2018) (cit. on p. 80).

[Ber+21] Jan Bouwe van den Berg, Maxime Breden, Jean-Philippe Lessard, and Lennaert van Veen. “Spontaneous Periodic Orbits in the Navier–Stokes Flow”. In: *Journal of Nonlinear Science* 31.2 (Mar. 2021), p. 41 (cit. on p. 10).

[BBS24] Jan Bouwe van den Berg, Maxime Breden, and Ray Sheombarsing. “Validated integration of semilinear parabolic PDEs”. In: *Numerische Mathematik* 156.4 (Aug. 2024), pp. 1219–1287 (cit. on pp. 10, 87).

[Ber14] Larbi Berrahmoune. “A variational approach to constrained controllability for distributed systems”. In: *Journal of Mathematical Analysis and Applications* 416.2 (2014). Publisher: Elsevier, pp. 805–823 (cit. on pp. 10, 87).

[Ber19] Larbi Berrahmoune. “Constrained null controllability for distributed systems and applications to hyperbolic-like equations”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 25 (2019). Publisher: EDP Sciences, p. 32 (cit. on pp. 10, 87).

[Ber20] Larbi Berrahmoune. “A variational approach to constrained null controllability for the heat equation”. In: *European Journal of Control* 52 (2020), pp. 42–48 (cit. on pp. 10, 87).

[BDM21] Viktor Bezborodov, Luca Di Persio, and Riccardo Muradore. “Minimal controllability time for systems with nonlinear drift under a compact convex state constraint”. In: *Automatica* 125 (2021), p. 109428 (cit. on p. 7).

[BT21] Loïc Bourdin and Emmanuel Trélat. “Robustness under control sampling of reachability in fixed time for nonlinear control systems”. In: *Mathematics of Control, Signals, and Systems* 33.3 (2021). Publisher: Springer, pp. 515–551 (cit. on p. 67).

[Boy13] Franck Boyer. “On the penalised HUM approach and its applications to the numerical approximation of null-controls for parabolic problems”. In: *ESAIM: Proceedings* 41 (Dec. 2013). Publisher: EDP Sciences, pp. 15–58 (cit. on p. 98).

[Boy22] Franck Boyer. “Controllability of linear parabolic equations and systems”. Master. France, Feb. 2022 (cit. on pp. 9, 86).

[Bra72] Robert F Brammer. “Controllability in linear autonomous systems with positive controllers”. In: *SIAM Journal on Control* 10.2 (1972). Publisher: SIAM, pp. 339–353 (cit. on pp. 7, 65, 86).

[BCL99] H. Brézis, P.G. Ciarlet, and J.L. Lions. *Analyse fonctionnelle: théorie et applications.* Collection Mathématiques appliquées pour la maîtrise. Dunod, 1999 (cit. on pp. 32, 34, 44).

[Bro08] Efim M Bronstein. “Approximation of convex sets by polytopes”. In: *Journal of Mathematical Sciences* 153.6 (2008), pp. 727–762 (cit. on pp. 117, 127, 128).

[CK22] Eduardo Casas and Karl Kunisch. “Boundary Control of Semilinear Parabolic Equations with Non-Smooth Point-wise-Integral Control Constraints in Time-Space”. In: *2022 American Control Conference (ACC).* 2022, pp. 284–289 (cit. on pp. 10, 87).

[CP11] Antonin Chambolle and Thomas Pock. “A first-order primal-dual algorithm for convex problems with applications to imaging”. In: *Journal of mathematical imaging and vision* 40.1 (2011). Publisher: Springer, pp. 120–145 (cit. on pp. 16, 26, 39–41, 76, 98, 99, 128).

[CF18] Dijian Chen and Kenji Fudjimoto. “Rendezvous Control of Spacecraft via Constrained Optimal Control Using Generating Functions”. In: *Transactions of the japan society for aeronautical and space sciences, aerospace technology Japan* 16.5 (2018), pp. 392–397 (cit. on p. 77).

[CR22] Mo Chen and Lionel Rosier. “Reachable states for the distributed control of the heat equation”. en. In: *Comptes Rendus. Mathématique* 360 (2022). Publisher: Académie des sciences, Paris, pp. 627–639 (cit. on pp. 9, 87, 103).

[CT18] Mo Chen and Claire J. Tomlin. “Hamilton–Jacobi Reachability: Some Recent Theoretical Advances and Applications in Unmanned Airspace Management”. In: *Annual Review of Control, Robotics, and Autonomous Systems* 1.1 (2018). _eprint: https://-doi.org/10.1146/annurev-control-060117-104941, pp. 333–358 (cit. on pp. 8, 66, 69, 87).

[CM24] Salah-Eddine Chorfi and Lahcen Maniar. “Controllability and Inverse Problems for Parabolic Systems with Dynamic Boundary Conditions”. In: *arXiv:2409.10302* (2024) (cit. on p. 9).

[Cro07] Michel Crouzeix. “Numerical range and functional calculus in Hilbert space”. In: *Journal of Functional Analysis* 244.2 (2007), pp. 668–690 (cit. on p. 34).

[CD03] Michel Crouzeix and Bernard Delyon. “Some estimates for analytic functions of strip or sectorial operators”. In: *Archiv der Mathematik* 81 (2003). Publisher: Springer, pp. 559–566 (cit. on pp. 34, 91, 94).

[CG19] Michel Crouzeix and Anne Greenbaum. “Spectral sets: numerical range and beyond”. In: *SIAM Journal on Matrix Analysis and Applications* 40.3 (2019). Publisher: SIAM, pp. 1087–1101 (cit. on pp. 34, 94).

[CP17] Michel Crouzeix and César Palencia. “The numerical range is a (1+2)-spectral set”. In: *SIAM Journal on Matrix Analysis and Applications* 38.2 (2017). Publisher: SIAM, pp. 649–655 (cit. on pp. 34, 94).

[DE18] Jérémi Dardé and Sylvain Ervedoza. “On the Reachable Set for the One-Dimensional Heat Equation”. In: *SIAM Journal on Control and Optimization* 56.3 (2018). _eprint: https://doi.org/10.1137/16M1093215, pp. 1692–1715 (cit. on pp. 9, 87).

[DZ18] Pighin Dario and Enrique Zuazua. “Controllability under positivity constraints of semilinear heat equations”. In: *Mathematical Control & Related Fields* 8.3&4 (2018), pp. 935–964 (cit. on p. 86).

[Day+04] Sarah Day, Yasuaki Hiraoka, Konstantin Mischaikow, and Toshiyuki Ogawa. “Rigorous Numerics for Global Dynamics: A Study of the Swift–Hohenberg Equation”. In: *SIAM Journal on Applied Dynamical Systems* 4 (Mar. 2004) (cit. on pp. 10, 87).

[De 01] Carl De Boor. *A practical guide to splines; rev. ed.* Applied mathematical sciences. Berlin: Springer, 2001 (cit. on pp. 51, 97).

[Egi+20] Michela Egidi, Ivica Nakić, Albrecht Seelmann, Matthias Täufer, Martin Tautenhahn, and Ivan Veselić. “Null-controllability and control cost estimates for the heat equation on unbounded and large bounded domains”. In: *Control theory of infinite-dimensional systems*. Springer. 2020, pp. 117–157 (cit. on p. 9).

[EN06] Klaus-Jochen Engel and Rainer Nagel. *A short course on operator semigroups*. Springer Science & Business Media, 2006 (cit. on p. 34).

[Erv20] Sylvain Ervedoza. “Control issues and linear projection constraints on the control and on the controlled trajectory”. English. In: *North-Western European Journal of Mathematics* 6 (2020), pp. 165–197 (cit. on pp. 9, 65).

[ELT22] Sylvain Ervedoza, Kevin Le Balc’h, and Marius Tucsnak. “Reachability results for perturbed heat equations”. In: *Journal of Functional Analysis* 283.10 (2022). Publisher: Elsevier, p. 109666 (cit. on pp. 9, 87).

[FHL92] M. Fashoro, O. Hajek, and K. Loparo. “Controllability properties of constrained linear systems”. In: *Journal of optimization theory and applications* 73 (1992). Publisher: Springer, pp. 329–346 (cit. on pp. 7, 65).

[Fis+15] Jaime F Fisac, Mo Chen, Claire J Tomlin, and S Shankar Sastry. “Reach-avoid problems with time-varying dynamics, targets and constraints”. In: *Proceedings of the 18th international conference on hybrid systems: computation and control*. 2015, pp. 11–20 (cit. on p. 8).

[GL08] Antoine Girard and Colas Le Guernic. “Efficient reachability analysis for linear systems using support functions”. In: *IFAC Proceedings Volumes* 41.2 (2008), pp. 8966–8971 (cit. on p. 8).

[GLM06] Antoine Girard, Colas Le Guernic, and Oded Maler. “Efficient computation of reachable sets of linear time-invariant systems with inputs”. In: *Hybrid Systems: Computation and Control: 9th International Workshop, HSCC 2006, Santa Barbara, CA, USA, March 29-31, 2006. Proceedings 9*. Springer, 2006, pp. 257–271 (cit. on pp. 8, 66).

[Góm19] Javier Gómez-Serrano. “Computer-assisted proofs in PDE: a survey”. In: *SeMA Journal* 76.3 (Sept. 2019), pp. 459–484 (cit. on pp. 10, 87).

[GP17] Eric Goubault and Sylvie Putot. “Forward inner-approximated reachability of non-linear continuous systems”. In: *Proceedings of the 20th international conference on hybrid systems: computation and control*. 2017, pp. 1–10 (cit. on p. 8).

[Gur+18] Thomas Gurriet, Andrew Singletary, Jacob Reher, Laurent Ciarletta, Eric Feron, and Aaron Ames. “Towards a Framework for Realizable Safety Critical Control through Active Set Invariance”. In: *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*. 2018, pp. 98–106 (cit. on p. 8).

[Haa06] M. Haase. *The Functional Calculus for Sectorial Operators*. Operator Theory: Advances and Applications. Birkhäuser Basel, 2006 (cit. on pp. 32–34).

[HK06] Zhi Han and Bruce H Krogh. “Reachability analysis of large-scale affine systems using low-dimensional polytopes”. In: *International Workshop on Hybrid Systems: Computation and Control*. Springer. 2006, pp. 287–301 (cit. on p. 8).

[HKT20] Andreas Hartmann, Karim Kellay, and Marius Tucsnak. “From the reachable space of the heat equation to Hilbert spaces of holomorphic functions”. In: *Journal of the European Mathematical Society* 22.10 (2020), pp. 3417–3440 (cit. on pp. 9, 87).

[Has+24] Ivan Hasenohr, Camille Pouchol, Yannick Privat, and Christophe Zhang. “Computer-assisted proofs of non-reachability for linear finite-dimensional control systems”. Mar. 2024 (cit. on pp. 15, 25, 64, 85).

[Hig08] Nicholas J. Higham. *Functions of matrices: theory and computation*. SIAM, 2008 (cit. on p. 74).

[HLQ02] Tingshu Hu, Zongli Lin, and Li Qiu. “An explicit description of null controllable regions of linear systems with saturating actuators”. In: *Systems & control letters* 47.1 (2002), pp. 65–78 (cit. on p. 7).

[Imm15] Fabian Immler. “Verified Reachability Analysis of Continuous Systems”. In: *Tools and Algorithms for the Construction and Analysis of Systems*. Ed. by Christel Baier and Cesare Tinelli. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 37–51 (cit. on pp. 8–10, 116).

[J-M07] J.-M. Coron. *Control and nonlinearity*. Vol. 136. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2007 (cit. on pp. 4, 6, 9, 65).

[Kap+21] Tomasz Kapela, Marian Mrozek, Daniel Wilczak, and Piotr Zgliczyński. “CAPD::DynSys: A flexible C++ toolbox for rigorous numerical analysis of dynamical systems”. In: *Communications in Nonlinear Science and Numerical Simulation* 101 (2021), p. 105578 (cit. on pp. 10, 87).

[KNT22] Karim Kellay, Thomas Normand, and Marius Tucsnak. “Sharp reachability results for the heat equation in one space dimension”. In: *Analysis & PDE* 15.4 (2022). Publisher: Mathematical Sciences Publishers, pp. 891–920 (cit. on pp. 9, 87).

[Kla96] Jerzy Klamka. “Constrained controllability of nonlinear systems”. In: *Journal of Mathematical Analysis and Applications* 201.2 (1996). Publisher: Elsevier, pp. 365–374 (cit. on pp. 9, 65).

[KA18] Shishir Kolathaya and Aaron D Ames. “Input-to-state safety with control barrier functions”. In: *IEEE control systems letters* 3.1 (2018), pp. 108–113 (cit. on p. 8).

[KBH18] H. Kong, E. Bartocci, and T. A. Henzinger. “Reachable Set Over-Approximation for Nonlinear Systems Using Piecewise Barrier Tubes”. In: *Computer Aided Verification*. Ed. by Hana Chockler and Georg Weissenbacher. Cham: Springer International Publishing, 2018, pp. 449–467 (cit. on pp. 8, 66, 87).

[KHJ13] Milan Korda, Didier Henrion, and Colin N. Jones. “Inner approximations of the region of attraction for polynomial dynamical systems”. In: *IFAC Proceedings Volumes* 46.23 (2013). 9th IFAC Symposium on Nonlinear Control Systems, pp. 534–539 (cit. on p. 8).

[Kra08] Mikhail I. Krastanov. “On the constrained small-time controllability of linear systems”. In: *Automatica* 44.9 (2008), pp. 2370–2374 (cit. on p. 7).

[KV02a] Alexander B Kurzhanski and Pravin Varaiya. “On ellipsoidal techniques for reachability analysis. Part I: external approximations”. In: *Optimization methods and software* 17.2 (2002). Publisher: Taylor & Francis, pp. 177–206 (cit. on pp. 8, 66).

[KV02b] Alexander B Kurzhanski and Pravin Varaiya. “On ellipsoidal techniques for reachability analysis. part ii: Internal approximations box-valued constraints”. In: *Optimization methods and software* 17.2 (2002). Publisher: Taylor & Francis, pp. 207–237 (cit. on pp. 8, 66).

[Le 09] Colas Le Guernic. “Reachability Analysis of Hybrid Systems with Linear Continuous Dynamics”. Theses. Université Joseph-Fourier - Grenoble I, Oct. 2009 (cit. on p. 8).

[LG09] Colas Le Guernic and Antoine Girard. “Reachability Analysis of Hybrid Systems Using Support Functions”. In: *Computer Aided Verification*. Ed. by Ahmed Bouajjani and Oded Maler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 540–554 (cit. on pp. 8, 66).

[LG10] Colas Le Guernic and Antoine Girard. “Reachability analysis of linear systems using support functions”. In: *Nonlinear Analysis: Hybrid Systems* 4.2 (May 2010). Publisher: Elsevier, pp. 250–262 (cit. on pp. 8, 66).

[LR94] G Lebeau and L Robbiano. “Contrôle exact de l’équation de la chaleur”. fre. In: *Séminaire Équations aux dérivées partielles (Polytechnique)* (1994). Publisher: Ecole Polytechnique, Centre de Mathématiques, pp. 1–11 (cit. on pp. 9, 86).

[LM86] E. B. Lee and L. Markus. *Foundations of optimal control theory*. Second. Robert E. Krieger Publishing Co., Inc., Melbourne, FL, 1986 (cit. on pp. 6, 14, 24, 69, 76, 83).

[LY12] Xungjing Li and Jiongmin Yong. *Optimal control theory for infinite dimensional systems*. Springer Science & Business Media, 2012 (cit. on pp. 10, 14, 24, 67).

[Lio88] Jacques-Louis Lions. “Exact controllability, stabilization and perturbations for distributed systems”. In: *SIAM review* 30.1 (1988), pp. 1–68 (cit. on p. 6).

[Lio92] Jacques-Louis Lions. “Remarks on approximate controllability”. In: *Journal d’Analyse Mathématique* 59.1 (1992). Publisher: Springer, p. 103 (cit. on pp. 6, 65).

[LM21] Pierre Lissy and Clément Moreau. “State-constrained controllability of linear reaction-diffusion systems”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 27 (2021). Publisher: EDP Sciences, p. 70 (cit. on pp. 9, 65).

[LTZ17] Jérôme Lohéac, Emmanuel Trélat, and Enrique Zuazua. “Minimal controllability time for the heat equation under unilateral state or control constraints”. In: *Mathematical Models and Methods in Applied Sciences* 27.09 (2017). Publisher: World Scientific, pp. 1587–1644 (cit. on pp. 7, 9).

[LTZ18] Jérôme Lohéac, Emmanuel Trélat, and Enrique Zuazua. “Minimal controllability time for finite-dimensional control systems under state constraints”. In: *Automatica* 96 (2018). Publisher: Elsevier, pp. 380–392 (cit. on pp. 7, 65, 86).

[LTZ21] Jérôme Lohéac, Emmanuel Trélat, and Enrique Zuazua. “Nonnegative control of finite-dimensional linear systems”. In: *Annales de l’Institut Henri Poincaré C, Analyse non linéaire* 38.2 (2021). Publisher: Elsevier, pp. 301–346 (cit. on pp. 7, 65, 86).

[MZ04] Sorin Micu and Enrique Zuazua. “An introduction to the controllability of partial differential equations”. In: *Quelques questions de théorie du contrôle. Sari, T., ed., Collection Travaux en Cours Hermann, to appear* (2004) (cit. on pp. 9, 86).

[MBT05] I.M. Mitchell, A.M. Bayen, and C.J. Tomlin. “A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games”. In: *IEEE Transactions on Automatic Control* 50.7 (2005), pp. 947–957 (cit. on pp. 8, 66, 69, 87).

[NPW19] Mitsuhiro T Nakao, Michael Plum, and Yoshitaka Watanabe. *Numerical verification methods and computer-assisted proofs for partial differential equations.* Vol. 53. Springer, 2019 (cit. on pp. 10, 87).

[Paz12] Amnon Pazy. *Semigroups of linear operators and applications to partial differential equations.* Vol. 44. Springer Science & Business Media, 2012 (cit. on p. 34).

[PN71] Thomas Pecsvaradi and Kumpati S. Narendra. “Reachable sets for linear dynamical systems”. In: *Information and Control* 19.4 (1971), pp. 319–344 (cit. on p. 8).

[PZ18] Dario Pighin and Enrique Zuazua. “Controllability under positivity constraints of semilinear heat equations”. In: *Mathematical Control and Related Fields* 8.3&4 (2018), pp. 935–964 (cit. on pp. 9, 65).

[PZ19] Dario Pighin and Enrique Zuazua. “Controllability under positivity constraints of multi-d wave equations”. In: *Trends in control theory and partial differential equations* (2019). Publisher: Springer, pp. 195–232 (cit. on pp. 9, 65).

[PTZ24] Camille Pouchol, Emmanuel Trélat, and Christophe Zhang. “Approximate control of parabolic equations with on-off shape controls by Fenchel duality”. In: *Annales de l’Institut Henri Poincaré C* (2024), pp. 1–43 (cit. on pp. 9, 65, 87).

[PTZ19] Camille Pouchol, Emmanuel Trélat, and Enrique Zuazua. “Phase portrait control for 1D monostable and bistable reaction–diffusion equations”. In: *Nonlinearity* 32.3 (2019). Publisher: IOP Publishing, p. 884 (cit. on p. 10).

[PJ04] Stephen Prajna and Ali Jadbabaie. “Safety Verification of Hybrid Systems Using Barrier Certificates”. In: *Hybrid Systems: Computation and Control.* Ed. by Rajeev Alur and George J. Pappas. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 477–492 (cit. on pp. 8, 66, 87).

[QSS06] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics.* Vol. 37. Springer Science & Business Media, 2006 (cit. on pp. 44, 46, 49, 73, 97).

[Res05] Jerzy Respondek. “Controllability of dynamical systems with constraints”. In: *Systems & Control Letters* 54.4 (2005). Publisher: Elsevier, pp. 293–314 (cit. on pp. 9, 65).

[Roc67] Ralph Rockafellar. “Duality and stability in extremum problems involving convex functions”. In: *Pacific Journal of Mathematics* 21.1 (1967). Publisher: Mathematical Sciences Publishers, pp. 167–187 (cit. on pp. 12, 22, 38).

[RW09] Ralph Rockafellar and Roger J-B Wets. *Variational analysis.* Vol. 317. Springer Science & Business Media, 2009 (cit. on p. 68).

[Roc70] Ralph Tyrell Rockafellar. “Convex Analysis”. In: (1970) (cit. on pp. 35, 42).

[Rud91] W. Rudin. *Functional Analysis.* International series in pure and applied mathematics. McGraw-Hill, 1991 (cit. on pp. 35, 36, 87).

[Rum99] Siegfried M Rump. “INTLAB—interval laboratory”. In: *Developments in reliable computing.* Springer, 1999, pp. 77–104 (cit. on pp. 10, 13, 23, 59, 65, 76, 89).

[SY71] Stephen H. Saperstone and James A. Yorke. “Controllability of linear oscillatory systems using positive controls”. In: *SIAM Journal on Control* 9.2 (1971), pp. 253–262 (cit. on p. 7).

[SL12] Heinz Schättler and Urszula Ledzewicz. *Geometric optimal control: theory, methods and examples.* Vol. 38. Springer, 2012 (cit. on p. 7).

[SB80] W.E. Schmitendorf and B.R. Barmish. “Null controllability of linear systems with constrained controls”. In: *SIAM Journal on control and optimization* 18.4 (1980), pp. 327–345 (cit. on p. 7).

[Sch13] Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory.* Vol. 151. Cambridge university press, 2013 (cit. on pp. 35, 42, 117).

[Ser20] Mohamed Serry. “Convergent under-approximations of reachable sets and tubes for linear uncertain systems”. In: *arXiv preprint arXiv:2002.04086* (2020) (cit. on p. 8).

[SV86] Nguyen Khoa Son and Nguyen Van Su. “Linear periodic control systems: controllability with restrained controls”. In: *Applied Mathematics and Optimization* 14.1 (1986), pp. 173–185 (cit. on p. 7).

[Tho07] Vidar Thomée. *Galerkin finite element methods for parabolic problems*. Vol. 25. Springer Science & Business Media, 2007 (cit. on pp. 47, 49).

[Tré23] Emmanuel Trélat. *Control in finite and infinite dimension*. _eprint: 2312.15925. Springer Singapore, 2023 (cit. on pp. 7, 9).

[Tuc11] Warwick Tucker. “Validated numerics: a short introduction to rigorous computations”. In: (2011) (cit. on p. 59).

[TW09] Marius Tucsnak and George Weiss. *Observation and control for operator semigroups*. Springer Science & Business Media, 2009 (cit. on pp. 6, 9, 88).

[Var00] Pravin Varaiya. “Reach Set Computation Using Optimal Control”. In: *Verification of Digital and Hybrid Systems*. Ed. by M. Kemal Inan and Robert P. Kurshan. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 323–331 (cit. on p. 8).

[Vel88] Vladimir Veliov. “On the controllability of control constrained linear systems”. In: *Mathematica Balkanica, New Series* 2.2–3 (1988), pp. 147–155 (cit. on p. 7).

[Vel92] Vladimir Veliov. “Second-order discrete approximation to linear differential inclusions”. In: *SIAM Journal on Numerical Analysis* 29.2 (1992), pp. 439–451 (cit. on p. 8).

[Wab+23] Kim P. Wabersich, Andrew J. Taylor, Jason J. Choi, Koushil Sreenath, Claire J. Tomlin, Aaron D. Ames, and Melanie N. Zeilinger. “Data-Driven Safety Filters: Hamilton-Jacobi Reachability, Control Barrier Functions, and Predictive Methods for Uncertain Systems”. In: *IEEE Control Systems Magazine* 43.5 (2023), pp. 137–177 (cit. on p. 8).

[Wan08] Gengsheng Wang. “L^∞-Null Controllability for the Heat Equation and Its Consequences for the Time Optimal Control Problem”. In: *Siam Journal on Control and Optimization - SIAM* 47 (Jan. 2008), pp. 1701–1720 (cit. on pp. 10, 87).

[WKA24] Mark Wetzlinger, Adrian Kulmburg, and Matthias Althoff. “Inner Approximations of Reachable Sets for Nonlinear Systems Using the Minkowski Difference”. In: *IEEE Control Systems Letters* (2024). Publisher: IEEE (cit. on p. 8).

[XFZ19] Bai Xue, Martin Fränzle, and Naijun Zhan. “Inner-approximating reachable sets for polynomial systems with time-varying uncertainties”. In: *IEEE Transactions on Automatic Control* 65.4 (2019), pp. 1468–1483 (cit. on p. 8).

[XSE16] Bai Xue, Zhikun She, and Arvind Easwaran. “Under-Approximating Backward Reachable Sets by Polytopes”. In: *Computer Aided Verification*. Ed. by Swarat Chaudhuri and Azadeh Farzan. Cham: Springer International Publishing, 2016, pp. 457–476 (cit. on p. 8).

[Zua10] Enrique Zuazua. “Switching control”. In: *Journal of the European Mathematical Society* 13.1 (2010), pp. 85–117 (cit. on pp. 9, 65).