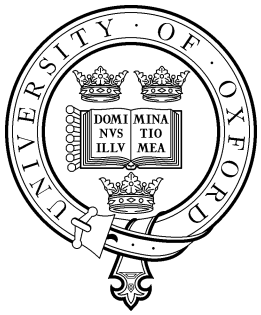
Decidability Boundaries in Linear Dynamical Systems



João Moreira de Sousa Pinto
St Cross College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2017

For their unconditional love and support,
I dedicate this thesis to my parents and to my brother.

Acknowledgements

First and foremost, I would like to thank my advisors: Professors Ben Worrell, Joël Ouaknine, and Elias Koutsoupias. I could not have hoped for better supervision during the past three years. This thesis would not exist without their guidance and encouragement.

I am also very grateful to all the wonderful people I have met whilst at Oxford. In particular, I would like to thank Andrew, Aris, Clara, Diogo, Edo, Francesco, Francisco, Gonçalo, Hosein, Hossam, Jasper, Jorge, Luis, Mafalda, Marcelo, Marina, Mathura, Miguel, Nicolas, Nikola, Ninad, Pascal, Philippos, Rajdeep, Raphael, Rita, Sebouh, Shahab, Smriti, Sophia, Stephen, Straulino, Tomás, and Ventsi. Their friendship is invaluable.

Moreover, I would like to thank St Cross College for the diverse and stimulating environment it promotes within its community, which were of great value to me. I will always remember college lunch very fondly.

It was an honour to have been examined by Professors Alessandro Abate (Confirmation of Status and Viva), Paul Goldberg (Transfer of Status), Angus MacIntyre (Confirmation of Status), Jonathan Pila (Transfer of Status), and Jeffrey Shallit (Viva).

Finally, I am obliged to acknowledge that, without the generous funding of the European Research Council (ALGAME grant) and of the Engineering and Physical Sciences Research Council, I would have been unable to conduct the research that led to this thesis.

Abstract

The object of this thesis is the study of the decidability properties of linear dynamical systems, which have fundamental ties to theoretical computer science, software verification, linear hybrid systems, and control theory.

In particular, we describe a method for deciding the termination of simple linear loops, partly solving a 10-year-old open problem of Tiwari [88] and Braverman [20]. We also study the membership problem for semigroups of matrix exponentials, which we show to be undecidable in general by reduction from Hilbert's Tenth Problem, and decidable for all instances where the matrices defining the semigroup commute. In turn, this entails the undecidability of the generalised versions of the Continuous Orbit and Skolem Problems to a multi-matrix setting. We also study point-to-point controllability for linear time-invariant systems, which is a central problem in control theory. For discrete-time systems, we show that this problem is undecidable when the set of controls is non-convex, and at least as hard as the Skolem Problem even when it is a convex polytope; for continuous-time systems, we show that this problem reduces to the Continuous Orbit Problem when the set of controls is a linear subspace, which entails decidability. Finally, we show how to decide whether all solutions of a given linear ordinary differential equation starting in a given convex polytope eventually leave it; this problem, which we call the "Polytope Escape Problem", relates to the liveness of states in linear hybrid automata.

Our results rely on a number of theorems from number theory, logic, and algebra, which we introduce in a self-contained way in the preamble to this thesis, together with a few new mathematical results of independent interest.

Statement of Originality

The contents of this thesis are partly based on my publications at the *ACM/SIAM Symposium on Discrete Algorithms 2015* [66], the *ACM/IEEE Symposium on Logic in Computer Science 2016* [65], and the *ACM International Conference on Hybrid Systems: Computation and Control 2017* [67]. At the time of writing, Chapter 6 and a number of sections of Chapter 4 were yet to be submitted for publication.

Contents

Chapter 1

Introduction

Dynamical systems have long been of interest to computer scientists. Many problems related to the behaviour of such systems, including the dynamics of polynomial differential equations, finitely generated matrix semigroups, or cellular automata, have been shown to be Turing-complete. However, despite the fact that some of these topics were studied over half a century ago, a surprising number of problems remain open, and we tackle some of them in this thesis.

In particular, we study dynamical systems with an algebraic nature, where the state space is typically a set of vectors or a set of matrices, and where the evolution of the system is determined by applying a *linear* operator to its state. A common property that is often studied is whether its state may ever reach a certain target. These systems are often described as *discrete-* or *continuous-time* depending on the nature of their evolution rule.

On the discrete-time front, *linear recurrence sequences* have been exhaustively studied. These correspond to sequences where each term is a fixed linear combination of the previous k terms (k is said to be the *order* of the linear recurrence sequence). The Skolem-Mahler-Lech theorem [85, 58, 54, 45] characterises the set of zeroes of such sequences as the union of finitely many arithmetic progressions and a finite set. Whilst it is known how to compute these arithmetic progressions [15], the general problem of determining whether a linear recurrence sequence ever hits zero, known as the *Skolem Problem*, remains open, and has been conjectured to be decidable. It has been shown that, when $k \leq 4$, the conjecture does indeed hold [90, in Russian], but the problem remains open for $k \geq 5$. Furthermore, other properties of these sequences, such as the Positivity Problem [10, 43, 53, 68, 69, 56] (which amounts to deciding whether a given linear recurrence sequence is always non-negative) and the Ultimate Positivity Problem [70], have been studied. In particular, these problems are known to be decidable when $k \leq 5$, and have shown to be hard when $k \geq 6$, in

the sense that their decidability would entail a substantial breakthrough in analytic number theory. Note that the Positivity Problem is at least as hard as Skolem's Problem [69], although the reduction causes a quadratic increase in the order of the sequence. Some of these results will be put to use in Chapter 3, and so we defer further details until then.

Kannan and Lipton's *Orbit Problem* [50], which consists in deciding whether an orbit of the form $(A^n \boldsymbol{x})_{n\in\mathbb{N}}$ ever hits a target point $\boldsymbol{y}$, was shown to be in **PTIME** in 1986. Further generalisations where the target is replaced by a small-dimension linear subspace [31] or a convex polytope [32] in a small-dimensional space have been shown to be decidable, but the general instances of these problems are, respectively, Skolem- and Positivity-hard.

As early as 1947, Markov showed in [60, in Russian] that the membership problem for finitely generated matrix sub-semigroups of $\mathbb{Z}^{6\times 6}$ is undecidable. In 1966, Mikhailova showed in [63, in Russian] that this problem is already undecidable for matrices in $SL(4,\mathbb{Z})$ (that is, 4×4 integer matrices of determinant 1). In 1970, Paterson established the undecidability of testing whether the zero matrix belongs to a given finitely generated sub-semigroup of 3×3 integer matrices [72]. The decidability of the membership problem for finitely generated semigroups of 2×2 integer matrices was only established in 2016, by Potapov and Semukhin in [75], and the case with *invertible* 3×3 integer matrices remains open. Moreover, the membership problem for finitely generated matrix semigroups of *commuting* integer matrices was shown to be decidable in any dimension by Babai, Beals, Cai, Ivanyos, and Luks in [4].

Another interesting problem relates to the study of matrix equations of the form

$$\prod_{i=1}^{k} A_i^{n_i} = C\,,$$

which were studied in [11], where it was shown to be undecidable whether they have any solution. This problem is of a very similar nature to the membership problem for finitely generated matrix semigroups, but here an order in the matrix products is enforced. Note that this undecidability result was established by reducing from Hilbert's Tenth Problem, as opposed to the Post Correspondence Problem, in similarity to what will frequently happen throughout this dissertation. In fact, we will need a strengthened version of this result where the matrices $A_1,\ldots,A_k,C$ are required to be invertible, as we shall see in Section 6.2.2.

On a different note, the study of *continous* models of computation started as early as 1941, when Shannon studied the General Purpose Analog Computer (GPAC). In

particular, he showed that the class of GPAC-computable functions corresponds to the set of functions which are components of a solution to an ordinary differential equation with a polynomial right-hand side [84]. Modulo technicalities, GPAC-computable functions have been shown to correspond to Turing-computable functions [17, 18], and even a neat characterisation of **PTIME** is known. Therefore, it is unsurprising that researchers have focused on linear ordinary differential equations, where numerous decidability results have been achieved.

The *Continuous Orbit Problem*, which amounts to deciding whether the unique solution of a given linear ordinary differential equation ever reaches a given target point, has been shown to be decidable in polynomial time [41, 28]. Some work has also been done on the *Continuous Skolem Problem*, which analogously to the Orbit Problem asks whether such a trajectory hits a given hyperplane (instead of a point). In [12], this problem was shown to be decidable when the ambient vector space has dimension 2. The dimension 3 case was shown to be decidable in [30], where it was also shown that, if Schanuel's conjecture is true, then all cases up to dimension 7 are decidable. The dimension 9 case was also shown to be hard in a number-theoretic sense; in particular, decidability would entail a major breakthrough in analytic number theory. Moreover, the version of this problem with a bounded time-horizon was shown to be decidable in [34] conditionally on Schanuel's conjecture being true.

Many other similar problems have been studied, and the reader should refer to the cited papers for further information, as well as the chapter-specific introductions.

Chapter 2

Mathematical Background

The objective of this chapter is threefold: to provide a concise overview of a number of mathematical results that will be put to use throughout this thesis; to acquaint the reader with a small amount of non-standard mathematical notation; to expose a few new mathematical results of independent interest.

2.1 Linear Algebra

2.1.1 Jordan Canonical Forms

Let $A \in \mathbb{Q}^{d\times d}$ be a square matrix with rational entries. The *minimal polynomial* of A is the unique monic polynomial $m(x) \in \mathbb{Q}[x]$ of least degree such that $m(A) = 0$. By the Cayley-Hamilton Theorem the degree of m is at most the dimension of A. The set $\sigma(A)$ of eigenvalues is the set of zeros of m, also known as the *spectrum* of A. The *index* of an eigenvalue λ, denoted by $\nu(\lambda)$, is its multiplicity as a zero of m. We use $\nu(A)$ to denote $\max_{\lambda\in\sigma(A)} \nu(\lambda)$: the maximum index over all eigenvalues of A. An eigenvalue λ is said to be *simple* if $\nu(\lambda) = 1$ and *repeated* otherwise. Given an eigenvalue $\lambda \in \sigma(A)$, we say that $\boldsymbol{v} \in \mathbb{C}^d$ is a *generalised eigenvector* of A if $\boldsymbol{v} \in \ker (A - \lambda I)^k$, for some $k \in \mathbb{N}$.

We denote the subspace of $\mathbb{C}^d$ spanned by the set of generalised eigenvectors associated with some eigenvalue λ by $\mathcal{V}_\lambda$. We denote the subspace of $\mathbb{C}^d$ spanned by the set of generalised eigenvectors associated with some real eigenvalue by $\mathcal{V}^r$. We likewise denote the subspace of $\mathbb{C}^d$ spanned by the set of generalised eigenvectors associated to eigenvalues with non-zero imaginary part by $\mathcal{V}^c$.

As a consequence of the existence of Jordan Canonical Forms, described later in

this subsection, each vector $\boldsymbol{v} \in \mathbb{C}^d$ can be written uniquely as

$$\boldsymbol{v} = \sum_{\lambda \in \sigma(A)} \boldsymbol{v}_\lambda, \tag{2.1}$$

where $\boldsymbol{v}_\lambda \in \mathcal{V}_\lambda$. It follows that $\boldsymbol{v}$ can also be uniquely written as $\boldsymbol{v} = \boldsymbol{v}^r + \boldsymbol{v}^c$, where $\boldsymbol{v}^r \in \mathcal{V}^r$ and $\boldsymbol{v}^c \in \mathcal{V}^c$.

We will need the following result:

Proposition 2.1. *Suppose that $\boldsymbol{v} \in \mathbb{R}^d$ and that $\boldsymbol{v} = \sum_{\lambda \in \sigma(A)} \boldsymbol{v}_\lambda$, where $\boldsymbol{v}_\lambda \in \mathcal{V}_\lambda$. Then $\boldsymbol{v}_{\overline{\lambda}}$ and $\boldsymbol{v}_\lambda$ are component-wise complex conjugates.*

Proof. Since A is real, $\boldsymbol{v}_\lambda \in \ker (A - \lambda I)^k$ implies that $\overline{\boldsymbol{v}_\lambda} \in \ker (A - \overline{\lambda} I)^k$ and hence that $\overline{\boldsymbol{v}_{\overline{\lambda}}} \in \mathcal{V}_\lambda$. The result follows from the fact that

$$\boldsymbol{0} = \boldsymbol{v} - \overline{\boldsymbol{v}} = \sum_{\lambda \in \sigma(A)} (\boldsymbol{v}_\lambda - \overline{\boldsymbol{v}_{\overline{\lambda}}})$$

and from uniqueness of the decomposition (2.1). □

We can write any matrix $A \in \mathbb{C}^{d \times d}$ as $A = Q^{-1} J Q$ for some invertible matrix Q and block diagonal Jordan matrix $J = \operatorname{diag}(J_1, \ldots, J_N)$, with each block J_i having the following form:

$$\begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{pmatrix}$$

Moreover, given a rational matrix A, its Jordan Normal Form $A = Q^{-1} J Q$ can be computed in polynomial time, as shown in [22].

Note that each vector $\boldsymbol{v}$ appearing as a column of the matrix Q^{-1} is a generalised eigenvector, and that the index $\nu(\lambda)$ of some eigenvalue λ corresponds to the dimension of the largest Jordan block associated with it.

One can obtain a closed-form expression for powers of block diagonal Jordan matrices, and use this to get a closed-form expression for the powers of any matrix A. In fact, if J_i is a $k \times k$ Jordan block associated with some eigenvalue λ, then

$$J_i^n = \begin{pmatrix} \lambda^n & n\lambda^{n-1} & \binom{n}{2}\lambda^{n-1} & \cdots & \binom{n}{k-1}\lambda^{n-k+1} \\ 0 & \lambda^n & n\lambda^{n-1} & \cdots & \binom{n}{k-2}\lambda^{n-k+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & n\lambda^{n-1} \\ 0 & 0 & 0 & \cdots & \lambda^n \end{pmatrix} \tag{2.2}$$

where $\binom{n}{j}$ is defined to be 0 when $n < j$.

2.1.2 Matrix exponentials

Given a matrix $A \in \mathbb{C}^{d\times d}$, its exponential is defined as

$$\exp(A) = \sum_{i=0}^{\infty} \frac{A^i}{i!}.$$

The series above always converges, and so the exponential of a matrix is always well defined. The standard way of computing $\exp(A)$ is by finding $P \in GL_d(\mathbb{C})$ such that $J = P^{-1}AP$ is in Jordan Canonical Form, and by using the fact that $\exp(A) = P\exp(J)P^{-1}$, where $\exp(J)$ is efficiently computable. When $A \in \overline{\mathbb{Q}}^{d\times d}$ (where $\overline{\mathbb{Q}}$ denotes the set of algebraic numbers; see Section 2.2.1), P can be taken to be in $GL_d(\overline{\mathbb{Q}})$; note that, due to Equation (2.2), if

$$J = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}$$

then

$$\exp(Jt) = \exp(\lambda t) \begin{pmatrix} 1 & t & \frac{t^2}{2} & \cdots & \frac{t^{k-1}}{(k-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{k-2}}{(k-2)!} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & t \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Then $\exp(J)$ can be obtained by setting $t = 1$, in particular $\exp(J)_{ij} = \frac{\exp(\lambda)}{(j-i)!}$ if $j \geq i$ and 0 otherwise.

When A and B commute, then so do $\exp(A)$ and $\exp(B)$. Moreover, when A and B have algebraic entries, the converse also holds, as shown in [91]. Also, when A and B commute, then $\exp(A)\exp(B) = \exp(A+B)$.

Proposition 2.2. *Let $\boldsymbol{v}$ lie in the generalised eigenspace $\mathcal{V}_\lambda$ for some $\lambda \in \sigma(A)$. Then $\boldsymbol{b}^T \exp(At)\boldsymbol{v}$ is a linear combination of terms of the form $t^n \exp(\lambda t)$.*

Proof. Note that, if $A = Q^{-1}JQ$ and $J = \mathrm{diag}(J_1, \ldots, J_N)$ is a block diagonal Jordan matrix, then $\exp(At) = Q^{-1}\exp(Jt)Q$ and $\exp(Jt) = \mathrm{diag}(\exp(J_1 t), \ldots, \exp(J_N t))$. The result follows by observing that $Q\boldsymbol{v}$ is zero in every component other than those pertaining the block corresponding to the eigenspace $\mathcal{V}_\lambda$. □

2.1.3 Matrix logarithms

The matrix B is said to be a logarithm of the matrix A if $\exp(B) = A$. It is well known that a logarithm of a matrix A exists if and only if A is invertible. However, matrix logarithms need not be unique. In fact, there exist matrices admitting uncountably many logarithms. See, for example, [38] and [47].

A matrix is said to be unitriangular if it is triangular and all its diagonal entries equal 1. Crucially, the following uniqueness result holds:

Theorem 2.3. *Given an upper unitriangular matrix $M \in \mathbb{C}^{d\times d}$, there exists a unique strictly upper triangular matrix L such that $\exp(L) = M$. Moreover, the entries of L lie in the number field $\mathbb{Q}(M_{i,j} : 1 \leq i, j \leq d)$.*

Proof. Firstly, we show that, for any strictly upper triangular matrix T and for any $1 < m < d$ and $i < j$, the term $(T^m)_{i,j}$ is a polynomial in the elements of the set $\{T_{r,s} : s - r < j - i\}$. This can be seen by induction on m, as each T^m is strictly upper triangular, and so

$$(T^m)_{i,j} = \sum_{l=1}^{d} (T^{m-1})_{i,l} T_{l,j} = \sum_{l=i+1}^{j-1} (T^{m-1})_{i,l} T_{l,j}.$$

Finally, we show, by induction on $j - i$, that each entry $L_{i,j}$ is a polynomial in the elements of the set

$$\{M_{i,j}\} \cup \{M_{r,s} : s - r < j - i\}.$$

If $j - i \leq 0$, then $L_{i,j} = 0$, so the claim holds. When $j - i > 0$, as L is nilpotent (as it is strictly upper triangular),

$$M_{i,j} = \exp(L)_{i,j} = L_{i,j} + \sum_{m=2}^{d-1} \frac{1}{m!} (L^m)_{i,j}$$

$$\Rightarrow L_{i,j} = M_{i,j} - \sum_{m=2}^{d-1} \frac{1}{m!} (L^m)_{i,j}.$$

The result now follows from the induction hypothesis and from our previous claim, as this argument can be used to both construct such a matrix L and to prove that it is uniquely determined. □

2.1.4 Properties of commuting matrices

We will now present a useful decomposition of $\mathbb{C}^d$ induced by the commuting matrices $A_1, \ldots, A_k \in \mathbb{C}^{d\times d}$. We remind the read that $\sigma(A_i)$ denotes the spectrum of the matrix A_i. In what follows, let

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k) \in \sigma(A_1) \times \cdots \times \sigma(A_k).$$

The generalised eigenspace of λ_i of A_i corresponds to $\ker (A_i - \lambda_i I)^d$, as the sequence $(\ker (A_i - \lambda_i I)^n)_{n\in\mathbb{N}}$ stabilises at most when $n = d$. With that in mind, we define the following subspaces of $\mathbb{C}^d$:

$$\mathcal{V}_{\boldsymbol{\lambda}} = \bigcap_{i=1}^{k} \ker (A_i - \lambda_i I)^d.$$

Also, let $\Sigma = \{\boldsymbol{\lambda} \in \sigma(A_1) \times \cdots \times \sigma(A_k) : \mathcal{V}_{\boldsymbol{\lambda}} \neq \{\mathbf{0}\}\}$. Below, $A_i \restriction_{\mathcal{V}_{\boldsymbol{\lambda}}}$ denotes the restriction of the linear operator A_i to the linear subspace $\mathcal{V}_{\boldsymbol{\lambda}}$, which is invariant under A_i.

Theorem 2.4. *For all $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k) \in \Sigma$ and for all $i \in \{1, \ldots, k\}$, the following properties hold:*

1. *$\mathcal{V}_{\boldsymbol{\lambda}}$ is invariant under A_i.*
2. *$\sigma(A_i \restriction_{\mathcal{V}_{\boldsymbol{\lambda}}}) = \{\lambda_i\}$.*
3. *$\mathbb{C}^d = \bigoplus_{\boldsymbol{\lambda}\in\Sigma} \mathcal{V}_{\boldsymbol{\lambda}}$.*

Proof. We show, by induction on k, that the subspaces $\mathcal{V}_{\boldsymbol{\lambda}}$ satisfy the properties above.

When $k = 1$, the result follows from the existence of Jordan Canonical Forms. When $k > 1$, suppose that $\sigma(A_k) = \{\mu_1, \ldots, \mu_m\}$, and let $\mathcal{U}_j = \ker (A_k - \mu_j I)^d$, for $j \in \{1, \ldots, m\}$. Again, it follows from the existence of Jordan Canonical Forms that

$$\mathbb{C}^d = \bigoplus_{j=1}^{m} \mathcal{U}_m.$$

In what follows, $i \in \{1, \ldots, k-1\}$ and $j \in \{1, \ldots, m\}$. Now, as A_k and A_i commute, so do $(A_k - \mu_j I)$ and A_i. Therefore, for all $\boldsymbol{v} \in \mathcal{U}_j$, $(A_k - \mu_j I)^d A_i \boldsymbol{v} = A_i (A - \mu_j I)^d \boldsymbol{v} = \mathbf{0}$, so $A_i \boldsymbol{v} \in \mathcal{U}_j$, that is, $\mathcal{U}_j$ is invariant under A_i. The result follows from applying the induction hypothesis to the commuting operators $A_i \restriction_{\mathcal{U}_j}$. □

We will also make use of the following well-known result on simultaneous triangularisation of commuting matrices. See, for example, [64].

Theorem 2.5. *Given k commuting matrices $A_1, \ldots, A_k \in \overline{\mathbb{Q}}^{d\times d}$, there exists a matrix $P \in GL_d(\overline{\mathbb{Q}})$ such that $P^{-1}A_iP$ is upper triangular for all $i \in \{1, \ldots, k\}$.*

2.2 Number Theory

2.2.1 Algebraic Number Theory

In this subsection, we introduce the main concepts in algebraic number theory necessary for understanding the hypothesis of the S-units theorem, stated below. In a later subsection, we shall justify the application of the aforementioned result in lower-bounding the dominant terms of linear recurrence sequences. We also briefly explain how one can effectively manipulate algebraic numbers. Should the reader be seeking an in-depth course in computational algebraic number theory, they can consult [35].

A complex number α is said to be *algebraic* if it is a zero of some non-zero polynomial with integer coefficients. Among those polynomials, there exists a unique one of minimal degree whose coefficients have no common factor and whose leading coefficient is positive, and it is said to be the *defining polynomial* of α, denoted by p_α, and it is always an irreducible polynomial. Moreover, if p_α is monic, α is said to be an *algebraic integer*. The degree of an algebraic number is defined as the degree of p_α, and its height as the maximum absolute value of the coefficients of p_α (also said to be the height of that polynomial). The zeros of p_α are said to be the *Galois conjugates* of α. Note that the complex conjugate of α is always a Galois conjugate of α. We denote the set of algebraic numbers by $\overline{\mathbb{Q}}$, and the set of algebraic integers by $\mathcal{O}$. For all $\alpha \in \overline{\mathbb{Q}}$, there exists some $n \in \mathbb{N}$ such that $n\alpha \in \mathcal{O}$. It is well known that $\overline{\mathbb{Q}}$ is a field and that $\mathcal{O}$ is a ring. A complex number that is not algebraic is said to be *transcendental*.

A *number field* of dimension d is a field extension K of $\mathbb{Q}$ whose degree as a vector-space over $\mathbb{Q}$ is d. In particular, $K \subseteq \overline{\mathbb{Q}}$ must hold. Recall that, in that case, there are exactly d monomorphisms $\sigma_i : K \to \mathbb{C}$ whose restriction over $\mathbb{Q}$ is the identity (and therefore these must map elements of K to their Galois conjugates). Given an algebraic number α, $\mathbb{Q}(\alpha)$ denotes the smallest number field containing α. Moreover, if $K = \mathbb{Q}(\alpha)$, then α is said to be a *primitive element* of K. It is well known that all number fields have a primitive element.

The *ring of integers* $\mathcal{O}_K$ of a number field K is the set of elements of K that are algebraic integers, that is, $\mathcal{O}_K = K \cap \mathcal{O}$. An ideal of $\mathcal{O}_K$ is an additive subgroup of $\mathcal{O}_K$ that is closed under multiplication by any element of $\mathcal{O}_K$. An ideal $\mathfrak{P}$ is said to be prime if $ab \in \mathfrak{P}$ implies $a \in \mathfrak{P}$ or $b \in \mathfrak{P}$. The following theorem is central in Algebraic Number Theory, and a proof can be found in [86]:

Theorem 2.6. *In any ring of integers, ideals can be uniquely factored as products of prime ideals up to permutation.*

The next theorem, by Evertse, van der Poorten, and Schlickewei, was established in [40, 89] to analyse the growth of linear recurrence sequences. It gives a very strong lower bound on the magnitude of sums of S-units, as defined below. Its key ingredient is Schlickewei's p-adic generalisation [83] of Schmidt's subspace theorem.

Let S be a finite set of prime ideals of the ring of integers $\mathcal{O}_K$ of a number field K. We say that $\alpha \in \mathcal{O}_K$ is an *S-unit* if all the ideals appearing in the prime factorisation of (α), that is, the principal ideal in $\mathcal{O}_K$ generated by α, are in S.

Theorem 2.7 (S-units)**.** *Let K be a number field, s be a positive integer, and S be a finite set of prime ideals of $\mathcal{O}_K$. Then for every $\varepsilon > 0$ there exists a constant C, depending only on s, K, S, and ε, with the following property. For every set of S-units $x_1, \ldots, x_s \in \mathcal{O}_K$ such that $\sum_{i \in I} x_i \neq 0$ for all non-empty $I \subseteq \{1, \ldots, s\}$,*

$$|x_1 + \cdots + x_s| \geq CYZ^{-\varepsilon}$$

where $Y = \max\{|x_j| : 1 \leq j \leq s\}$ and $Z = \max\{\sigma_i(x_j) : 1 \leq j \leq s, 1 \leq i \leq d\}$ and σ_i represent the different monomorphisms from K to $\mathbb{C}$.

2.2.1.1 Manipulating algebraic numbers

The following separation bound allows us to effectively represent an arbitrary algebraic number by keeping its defining polynomial, a sufficiently accurate estimate for the zero we want to represent, and an upper bound on the error. We call this its *standard/canonical representation*.

Lemma 2.8 (Mignotte)**.** *Let $f \in \mathbb{Z}[x]$. Then*

$$f(\alpha_1) = 0 = f(\alpha_2) \Rightarrow |\alpha_1 - \alpha_2| > \frac{\sqrt{6}}{d^{(d+1)/2} H^{d-1}} \tag{2.3}$$

where d and H are respectively the degree and height of f.

It is well known that arithmetic operations and equality testing on these caninocal representations can be done in time polynomial on their size, since one can

- compute polynomially many bits of the zeros of any polynomial $p \in \mathbb{Q}[x]$ in polynomial time, due to the work of Pan in [71];
- find the defining polynomial of an algebraic number by factoring the polynomial in its description in polynomial time using the LLL algorithm [55];
- use the sub-resultant algorithm (see Algorithm 3.3.7 in [35]) and the two aforementioned procedures to compute canonical representations of sums, differences, multiplications, and divisions of canonically represented algebraic numbers.

Moreover, we need to know how to decide whether a given canonically represented algebraic number α is a root of unity, that is, whether $\alpha^r = 1$ for some r. If that is the case, then its defining polynomial will be the r-th cyclotomic polynomial, which has degree $\phi(r)$, if r is taken to be minimal, that is, if α is a primitive r-th root of unity. The following (crude) lower bound on $\phi(r)$ allows us to decide this problem in polynomial time, assuming that the degree of α is given in unary. It follows from the obvious fact that $\phi(r) \geq \pi(r)$, where $\pi(r)$ denotes the number of prime numbers in the set $\{1, \ldots, r\}$, and from the fact that $\pi(r) > r/\log(r) > \sqrt{r/2}$ for $r \geq 17$ [78, Corollary 1]. It can be readily verified that the result below also holds for $1 \leq r \leq 16$. Note that much better bounds are known; for example, $\phi(r) = \Omega(r/\log\log(r))$ [78, Theorem 15].

Lemma 2.9. *Let ϕ be Euler's totient function. Then $\phi(r) \geq \sqrt{r/2}$. Therefore, if α has degree n and is a r'th root of unity, then $r \leq 2n^2$.*

Therefore, in order to decide whether an algebraic number α of degree n is a root of unity, we check whether it is a r-th root of unity, for each $r \leq 2n^2$. In order to test whether α is a r-th root of unity, it suffices to see whether $\gcd(p_\alpha, x^r - 1) = p_\alpha$, since we know that $x^r - 1$ is the product of each d-th cyclotomic polynomial, with d ranging over the divisors of r.

2.2.1.2 Integral solutions of linear equations with algebraic coefficients

Consider a non-zero matrix $K \in \overline{\mathbb{Q}}^{r \times d}$ and vector $\boldsymbol{k} \in \overline{\mathbb{Q}}^r$. The following proposition shows how to compute a representation of the set $\{\boldsymbol{x} \in \mathbb{Z}^d : K\boldsymbol{x} = \boldsymbol{k}\}$.

Proposition 2.10. *Let $S = \{\boldsymbol{x} \in \mathbb{Z}^d : K\boldsymbol{x} = \boldsymbol{k}\}$. If $S \neq \emptyset$, then there exist $\boldsymbol{x}_0 \in \mathbb{Z}^d$ and $M \in \mathbb{Z}^{d \times s}$ such that $S = \boldsymbol{x}_0 + \{M\boldsymbol{y} : \boldsymbol{y} \in \mathbb{Z}^s\}$.*

Proof. Let θ denote a primitive element of the number field generated by the entries of K and $\boldsymbol{k}$. Let the degree of this extension, which equals the degree of θ, be D. Then for $\boldsymbol{x} \in \mathbb{Z}^d$ one can write

$$\begin{aligned} K\boldsymbol{x} = \boldsymbol{k} &\Leftrightarrow \left(\sum_{i=0}^{D-1} N_i \theta^i\right) \boldsymbol{x} = \sum_{i=0}^{D-1} \boldsymbol{k}_i \theta^i \\ &\Leftrightarrow N_i \boldsymbol{x} = \boldsymbol{k}_i, \forall i \in \{0, \ldots, D-1\}, \end{aligned}$$

for some integer matrices $N_0, \ldots, N_{D-1} \in \mathbb{Z}^{r \times d}$ and integer vectors $\boldsymbol{k}_0, \ldots, \boldsymbol{k}_{D-1} \in \mathbb{Z}^r$. We take $\boldsymbol{x}_0$ to be any solution of this system, and select the columns of M to be a minimal set generating

$$\mathcal{G} = \{\boldsymbol{x} \in \mathbb{Z}^d : \forall i \in \{0, \ldots, D-1\}, N_i \boldsymbol{x} = \boldsymbol{0}\}.$$

Note that, since $\mathcal{G}$ is a subgroup of the finitely generated abelian group $\mathbb{Z}^d$, $\mathcal{G}$ itself must be finitely generated. □

2.2.2 Transcendental Number Theory

A number of the form

$$\alpha_0 + \alpha_1 \log(\beta_1) + \cdots + \alpha_n \log(\beta_n),$$

where $\alpha_0, \ldots, \alpha_n, \beta_1, \ldots, \beta_n$ are algebraic numbers, is said to be a *linear form in logarithms of algebraic numbers*. Note that the set of linear forms in logarithms of algebraic numbers is closed under addition and under multiplication by algebraic numbers, as well as under complex conjugation.

The following result, together with Theorem 2.23, yields a method for comparing linear forms in logarithms of algebraic numbers. Note that there are other theorems by Alan Baker that would allow us to do this directly, namely by providing lower bounds on the absolute value of non-zero linear forms in logarithms of algebraic numbers as a function of the degrees and heights of the defining polynomials of the algebraic numbers appearing therein; for simplicity, we will not discuss these results in this thesis. For a proof, see [6] and [7].

Theorem 2.11 (Baker)**.** *Let* $\alpha_1, \ldots, \alpha_m \in \overline{\mathbb{Q}} \setminus \{0\}$*. If*

$$\log(\alpha_1), \ldots, \log(\alpha_m)$$

are linearly independent over $\mathbb{Q}$*, then*

$$1, \log(\alpha_1), \ldots, \log(\alpha_m)$$

are linearly independent over $\overline{\mathbb{Q}}$*.*

The theorem below was proved by Ferdinand von Lindemann in 1882, and later generalised by Karl Weierstrass in what is now known as the Lindemann-Weierstrass theorem. As a historical note, this result marked the first proof of transcendence of π, which immediately follows from it.

Theorem 2.12 (Lindemann). *If $\alpha \in \overline{\mathbb{Q}} \setminus \{0\}$, then e^{α} is transcendental.*

We will later present a result that holds if the following conjecture, which generalises many other theorems from transcendental number theorem (including Baker's theorem and the Lindemann-Weierstrass theorem), is true.

Conjecture 2.13 (Schanuel). *If $\alpha_1, \ldots, \alpha_m \in \mathbb{C}$ are linearly independent over $\mathbb{Q}$, then $\mathbb{Q}(\alpha_1, \exp(\alpha_1), \ldots, \alpha_m, \exp(\alpha_m))$ has transcendence degree at least m.*

Note that the transcendence degree of a field extension of $\mathbb{Q}$ is the cardinality of the largest algebraically independent subset thereof, so having transcendence degree at least m means that there is a set of at least m elements that satisfies no non-zero polynomial relation with integer coefficients. Whilst we will not make use of this conjecture directly, we will use Theorem 2.21, which does rely on Schanuel's conjecture being true. Obviously, all conditional results will be marked as such.

2.2.3 Diophantine Approximation

The following result, due to Leopold Kronecker, on simultaneous Diophantine approximation, generalises Dirichlet's Approximation Theorem. We denote the *group of additive relations* of $\boldsymbol{v} \in \mathbb{C}^d$ by

$$\mathcal{A}(\boldsymbol{v}) = \{\boldsymbol{z} \in \mathbb{Z}^d : \boldsymbol{z} \cdot \boldsymbol{v} \in \mathbb{Z}\}.$$

Throughout this thesis, dist refers to the l_1 distance.

Theorem 2.14 (Kronecker). *Let $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha_k} \in \mathbb{R}^d$ and $\boldsymbol{\beta} \in \mathbb{R}^d$. The following are equivalent:*

1. *For any $\varepsilon > 0$, there exists $\boldsymbol{n} \in \mathbb{N}^k$ such that*

$$\operatorname{dist}(\boldsymbol{\beta} + \sum_{i=1}^{k} n_i \boldsymbol{\alpha}_i, \mathbb{Z}^d) \leq \varepsilon.$$

2. *The following relation holds:*

$$\bigcap_{i=1}^{k} \mathcal{A}(\boldsymbol{\alpha}_i) \subseteq \mathcal{A}(\boldsymbol{\beta}).$$

A proof of this result can be found in [25]. Note that the second condition essentially states that all the integer relations that are satisfied by all the $\boldsymbol{\alpha}_i$ are also satisfied by $\boldsymbol{\beta}$.

We will also need the following result, which is a weak form of [51, Corollary 2.8].

Theorem 2.15. *Suppose that $\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k\} \subseteq \mathbb{R}^d$ and that $\mathcal{C}^\perp \cap \mathbb{Z}^d = \{\boldsymbol{0}\}$. Then for any $\boldsymbol{q} \in \mathbb{R}^d$ and for any $\varepsilon > 0$ there exist non-negative reals $\lambda_1, \ldots, \lambda_k$ such that*

$$\operatorname{dist}(\boldsymbol{q} + \sum_{i=1}^{k} \lambda_i \boldsymbol{c}_i, \mathbb{Z}^d) \leq \varepsilon \,.$$

In order to compare Theorem 2.15 to Theorem 2.14, note that

$$\bigcap_{i=1}^{k} \mathcal{A}(\boldsymbol{c}_i) = \{\boldsymbol{0}\} \Rightarrow \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k\}^\perp \cap \mathbb{Z}^d = \{\boldsymbol{0}\} \,.$$

2.3 Logic

2.3.1 First-Order Theory of the Reals

Let $\boldsymbol{x} = (x_1, \ldots, x_m)$ be a vector of m real-valued variables, and let $\sigma(\boldsymbol{x})$ be a Boolean combination of atomic predicates of the form $g(\boldsymbol{x}) \sim 0$, where each $g(\boldsymbol{x})$ is a polynomial with integer coefficients in the variables $\boldsymbol{x}$, and $\sim$ is either $>$ or $=$. The set of true sentences of the form $\phi = Q_1 x_1 \cdots Q_m x_m \sigma(\boldsymbol{x})$, where Q_i is either $\exists$ or $\forall$, is called the *first-order theory of the reals*.

A set $S \subseteq \mathbb{R}^n$ is said to be *semi-algebraic* if it is a Boolean combination of sets of the form $\{\boldsymbol{x} \in \mathbb{R}^n : p(\boldsymbol{x}) \geq 0\}$, where p is a polynomial with integer coefficients. Equivalently, the semi-algebraic sets are those definable by the quantifier-free first-order formulas over the structure $(\mathbb{R}, <, +, \cdot, 0, 1)$.

Theorem 2.16 (Tarski-Seidenberg)**.** *The first-order theory of the reals admits a constructive method for quantifier elimination. In particular, it is a decidable theory. [87]*

From Theorem 2.16, it follows that the semi-algebraic sets are precisely the first-order definable sets (that is, the use of quantifiers does not augment the class of semi-algebraic sets).

We also remark that our standard representation of algebraic numbers, described in Section 2.2.1.1, allows us to write them explicitly in the first-order theory of the reals, that is, given $\alpha \in \overline{\mathbb{Q}}$, there exists a sentence $\sigma(x)$ such that $\sigma(x)$ is true if

and only if $x = \alpha$. Thus, we allow their use when defining semi-algebraic sets, for simplicity.

The complexity class $\exists\mathbb{R}$ is defined as the set of problems having a polynomial-time many-one reduction to the existential theory of the reals. It was shown in [24] that $\exists\mathbb{R} \subseteq \mathbf{PSPACE}$.

The reader seeking a comprehensive study of the computational complexity of quantifier elimination may refer to [77] and [9]. We shall make use of the following result by Basu, Pollack, and Roy [8], which provides an upper bound on the space complexity of quantifier elimination:

Theorem 2.17. *Given a set $\mathcal{Q} = \{q_1, \ldots, q_s\}$ of s polynomials each of degree at most D, in $h + d$ variables, and a first-order formula*

$$\Phi(\boldsymbol{x}) = Qy_1 \ldots Qy_h F(q_1(\boldsymbol{x}, \boldsymbol{y}), \ldots, q_s(\boldsymbol{x}, \boldsymbol{y}))\,,$$

where $Q \in \{\exists, \forall\}$, F is a quantifier-free Boolean combination with atomic elements of the form $q_i(\boldsymbol{x}, \boldsymbol{y}) \sim 0$, where $\sim \in \{>, =\}$, there exists a quantifier-free formula

$$\Psi(\boldsymbol{x}) = \bigwedge_{i=1}^{I} \bigvee_{j=1}^{J_i} q_{ij}(\boldsymbol{x}) \sim 0\,,$$

where $I \leq (sD)^{O(hd)}$, each $J_i \leq (sD)^{O(d)}$, the degrees of the polynomials q_{ij} are bounded by D^d, and the bit-sizes of the heights of the polynomials in the quantifier-free formula are only polynomially larger than those of $q_1, \ldots, q_s$.

We also make use of the following lemmas:

Lemma 2.18. *If $X \subseteq \mathbb{R}^d$ is semi-algebraic and non-empty, then $X \cap \overline{\mathbb{Q}}^d \neq \emptyset$.*

Proof. We prove this result by strong induction on d. Since X is semi-algebraic, there exists a quantifier-free sentence in the first-order theory of the reals σ such that $X = \{x \in \mathbb{R}^d \mid \sigma(x)\}$.

Suppose that $d > 1$. Letting $X_1 = \{x_d \in \mathbb{R} \mid \exists x_1, \ldots, x_{d-1} \in \mathbb{R}^{d-1}, \sigma(x_1, \ldots, x_d)\}$ and since $X_1 \neq \emptyset$ is semi-algebraic, by the induction hypothesis, there must be $x_d^* \in \overline{\mathbb{Q}} \cap X_1$. Moreover, we can define $X_2 = \{(x_2, \ldots, x_d) \in \mathbb{R}^{d-1} \mid \sigma(x_1^*, x_2, \ldots, x_n)\}$, which is non-empty and semi-algebraic, and again by the induction hypothesis there exists some $(x_2^*, \ldots, x_d^*) \in \overline{\mathbb{Q}}^{d-1} \cap X_2$.

It remains to prove this statement for $d = 1$. In that case, X must be a finite union of intervals and points. Clearly $\overline{\mathbb{Q}}$ is dense in any interval, and each of these isolated points x corresponds to some constraint $g(x) = 0$, which implies that x must be algebraic, since g has integer coefficients. □

Lemma 2.19. *If $X \subseteq \mathbb{R}^d$ is semi-algebraic, then $X \cap \overline{\mathbb{Q}}^d$ is dense in X.*

Proof. Pick $x \in X$ and $\varepsilon > 0$ arbitrarily. Let $y \in \mathbb{Q}^d$ be such that $\|x - y\| < \varepsilon/2$. Since $B(y, \varepsilon/2)$ is semi-algebraic, so must be $X \cap B(y, \varepsilon/2)$, and so this set must contain an algebraic point, since it is nonempty (x is in it), and that point must therefore be at distance at most ε from x, by the triangle inequality. □

Lemma 2.20. *If $X \subseteq \mathbb{R}^d$ is semi-algebraic, then so is* $\mathrm{Cl}(X)$[1].

Proof. Let σ be a sentence in the first-order theory of the reals such that $X = \{x \in \mathbb{R}^d \mid \sigma(x)\}$. Whence

$$\mathrm{Cl}(X) = \{x \in \mathbb{R}^d \mid \forall \varepsilon > 0, \exists y \in \mathbb{R}^d, \sigma(y) \wedge y \in B(x, \varepsilon)\}.$$

□

Finally, we will need the following theorem, shown by MacIntyre and Wilkie [57], concerning the decidability of the extension of the first-order theory of the reals with real exponentiation and bounded sin and cos functions.

Theorem 2.21 (Wilkie and MacIntyre)**.** *If Schanuel's conjecture is true, then, for each $n \in \mathbb{N}$, $FO(\mathbb{R}, +, \cdot, <, =, \exp\restriction_{\mathbb{R}}, \cos\restriction_{[0,n]}, \sin\restriction_{[0,n]})$ is decidable.*

2.3.2 Hilbert's Tenth Problem

A set $S \subseteq \mathbb{Z}^d$ is called *Diophantine* if there exist $m \in \mathbb{N}$ and $p \in \mathbb{Z}[x_1, \ldots, x_{d+m}]$ such that

$$S = \{\boldsymbol{x} \in \mathbb{Z}^d : \exists \boldsymbol{y} \in \mathbb{Z}^m,\, p(\boldsymbol{x}, \boldsymbol{y}) = 0\}\,.$$

Equivalently, Diophantine sets correspond to those that are definable in the existential branch of the first-order theory of the ring of integers.

In his famous list of 23 problems, Hilbert asked whether there exists an algorithm for deciding if a given polynomial $p \in \mathbb{Z}[x_1, \ldots, x_d]$ admits any zero $\boldsymbol{x} \in \mathbb{Z}^d$; this was the tenth problem in his list.

The following celebrated theorem, due to Yuri Matiyasevich, settled this question negatively, namely by showing that it is an undecidable problem; see [62] for a self-contained proof.

Theorem 2.22 (Matiyasevich)**.** *The recursively enumerable subsets of $\mathbb{Z}^d$ are Diophantine.*

[1] $\mathrm{Cl}(X)$ denotes the topological closure of X.

It is obvious that the converse is also true, namely that all Diophantine sets are recursively enumerable, which follows from the existence of a computable surjection mapping $\mathbb{N}$ to $\mathbb{Z}^d$.

Theorem 2.22 can be seen as a strengthening of the famous Gödel-Rosser incompleteness theorem, which states that the first-order theory of the ring of integers is undecidable. Namely, this result shows that the existential branch of that theory is already undecidable.

2.4 Groups of Multiplicative Relations

This section introduces some concepts concerning groups of multiplicative relations among algebraic numbers. We assume a natural first-order interpretation of the field of complex numbers in the ordered field of real numbers (in which each complex number is encoded as a pair comprising its real and imaginary parts). Under this interpretation we refer to sets of complex numbers as being semi-algebraic and first-order definable.

Let $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$. We define the *s-dimensional torus* to be $\mathbb{T}^s$, considered as a group under componentwise multiplication. Then the function $x \mapsto \exp(2\pi i x)$ is a homomorphism from the additive group of real numbers to $\mathbb{T}$, with kernel the subgroup of integers. By abuse of notation, we will also allow exp to be applied component-wise to a vector of reals.

Given a tuple of algebraic numbers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s) \in \mathbb{T}^s$, we consider how to effectively represent the *orbit*

$$\{\boldsymbol{\lambda}^n : n \in \mathbb{N}\}.$$

More precisely, we will give an algebraic representation of the topological closure of that orbit in $\mathbb{T}^s$.

The *group of multiplicative relations* of $\boldsymbol{\lambda}$, which is an additive subgroup of $\mathbb{Z}^s$, is defined as

$$\mathcal{L}(\boldsymbol{\lambda}) = \{\boldsymbol{v} \in \mathbb{Z}^s : \boldsymbol{\lambda}^{\boldsymbol{v}} = 1\},$$

where $\boldsymbol{\lambda}^{\boldsymbol{v}}$ is defined to be $\lambda_1^{v_1} \cdots \lambda_s^{v_s}$ for $\boldsymbol{v} \in \mathbb{Z}^s$, that is, exponentiation acts coordinatewise.

Since $\mathbb{Z}^s$ is a free abelian group, its subgroups are also free. In particular, $\mathcal{L}(\boldsymbol{\lambda})$ has a finite basis. The following powerful theorem of Masser [61] gives bounds on the magnitude of the components of such a basis.

Note that $\log(\alpha_1), \ldots, \log(\alpha_m)$ are linearly independent over $\mathbb{Q}$ if and only if

$$\mathcal{L}(\alpha_1, \ldots, \alpha_m) = \{\mathbf{0}\}.$$

Together with Theorem 2.11, Masser's theorem allows us to eliminate all algebraic relations in the description of linear forms in logarithms of algebraic numbers, and therefore also to test whether a linear form in logarithms of algebraic numbers is zero.

Theorem 2.23 (Masser)**.** *The free abelian group $\mathcal{L}(\boldsymbol{\lambda})$ has a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l \in \mathbb{Z}^s$ for which*

$$\max_{1 \leq i \leq l, 1 \leq j \leq s} |v_{i,j}| \leq (D \log H)^{O(s^2)}$$

where H and D bound respectively the heights and degrees of all the λ_i.

Membership of a tuple $\boldsymbol{v} \in \mathbb{Z}^s$ in $\mathcal{L}(\boldsymbol{\lambda})$ can be computed in polynomial space, using a decision procedure for the existential theory of the reals. In combination with Theorem 2.23, it follows that we can compute a basis for $\mathcal{L}(\boldsymbol{\lambda})$ in polynomial space by brute-force search (due to Savitch's theorem [82]).

Corresponding to $\mathcal{L}(\boldsymbol{\lambda})$, we consider the following multiplicative subgroup of $\mathbb{T}^s$:

$$T(\boldsymbol{\lambda}) = \{\boldsymbol{\mu} \in \mathbb{T}^s : \forall \boldsymbol{v} \in \mathcal{L}(\boldsymbol{\lambda}),\ \boldsymbol{\mu}^{\boldsymbol{v}} = 1\}\,.$$

If $\mathcal{B}$ is a basis of $\mathcal{L}(\boldsymbol{\lambda})$ then we can equivalently characterise $T(\boldsymbol{\lambda})$ as

$$\{\boldsymbol{\mu} \in \mathbb{T}^s : \forall \boldsymbol{v} \in \mathcal{B},\ \boldsymbol{\mu}^{\boldsymbol{v}} = 1\}\,.$$

Crucially, this finitary characterisation allows us to represent $T(\boldsymbol{\lambda})$ as a semi-algebraic set.

We will use Theorem 2.14 to show that the orbit $\{\boldsymbol{\lambda}^n : n \in \mathbb{N}\}$ is a dense subset of $T(\boldsymbol{\lambda})$.

Theorem 2.24. *Let $\boldsymbol{\lambda} \in \mathbb{T}^s$. Then the orbit $\{\boldsymbol{\lambda}^n : n \in \mathbb{N}\}$ is a dense subset of $T(\boldsymbol{\lambda})$.*

Proof. Let $\boldsymbol{\mu}$ be an arbitrary element of $T(\boldsymbol{\lambda})$, and let $\boldsymbol{\theta} \in \mathbb{R}^s$ be such that $\boldsymbol{\lambda} = \exp(2\pi i \boldsymbol{\theta})$ (with exponentiation operating coordinatewise). Notice that $\boldsymbol{\lambda}^{\boldsymbol{v}} = 1$ if and only if $\boldsymbol{v}^T \boldsymbol{\theta} \in \mathbb{Z}$. Similarly, we define $\boldsymbol{\psi} \in \mathbb{R}^s$ to be such that $\boldsymbol{\mu} = \exp(2\pi i \boldsymbol{\psi})$. Then the second condition in the statement of Theorem 2.14 holds for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. Thus, given $\varepsilon > 0$, there exist a non-negative integer n and a vector $\boldsymbol{p} \in \mathbb{Z}^s$ such that $\operatorname{dist}(n\boldsymbol{\theta} - \boldsymbol{\psi}, \boldsymbol{p}) \leq \varepsilon$. Whence

$$\|\boldsymbol{\lambda}^n - \boldsymbol{\mu}\|_\infty = \|\exp(2\pi i(n\boldsymbol{\theta} - \boldsymbol{p})) - \exp(2\pi i \boldsymbol{\psi})\|_\infty \leq \|2\pi(n\boldsymbol{\theta} - \boldsymbol{p} - \boldsymbol{\psi})\|_\infty \leq 2\pi\varepsilon\,.$$

Given that ε was arbitrary, it follows that $\{\boldsymbol{\lambda}^n : n \in \mathbb{N}\}$ is dense in $T(\boldsymbol{\lambda})$. □

We will also need the following simple corollary:

Corollary 2.25. *Let* $\boldsymbol{\theta} \in \mathbb{R}^s$ *be such that* $\mathcal{A}(\boldsymbol{\theta}) = \{\mathbf{0}\}$. *Then* $\{\exp(2\pi i n\boldsymbol{\theta}) : n \in \mathbb{N}\}$ *is a dense subset of* $\mathbb{T}^s$.

Proof. This result follows from the fact that

$$\mathcal{L}(\exp(2\pi i\boldsymbol{\theta})) = \mathcal{A}(\boldsymbol{\theta}) = \{\mathbf{0}\}\,.$$

Therefore, $T(\boldsymbol{\lambda}) = \mathbb{T}^s$, which establishes the result. □

2.5 Linear Recurrence Sequences

A *linear recurrence sequence* of order k over $\mathbb{Q}$ is a sequence $(u_n)_{n\in\mathbb{N}} \subseteq \mathbb{Q}$ with the property that there exist coefficients $\alpha_1, \ldots, \alpha_k \in \mathbb{Q}$ such that, for all $n \geq k$,

$$u_n = \sum_{i=1}^{k} \alpha_i u_{n-i}\,.$$

Equivalently, due to the Cayley-Hamilton theorem, linear recurrence sequences of order k correspond to sequences of the form $u_n = \boldsymbol{b}^T A^n \boldsymbol{x}$. When the matrix A is diagonalisable, the linear recurrence $(u_n)_{n\in\mathbb{N}}$ is said to be *simple*. The *characteristic roots* of u correspond to the eigenvalues of A.

It is well-known that linear recurrence sequences correspond to sequences admitting a representation of the form

$$u_n = \sum_{i=1}^{k} p_i(n)\lambda_i^n \tag{2.4}$$

where $\lambda_1, \ldots, \lambda_k$ are the characteristic roots of u, as can easily be shown by considering the Jordan canonical form of the matricial form of this recurrence.

Moreover, the sequences admitting a representation of the form described in Equation (2.4), where the polynomials p_i are constant, correspond to simple linear recurrence sequences.

Finally, if $(u_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$, then the polynomials p_i and constants λ_i in Equation (2.4) satisfy the relation

$$\lambda_i = \overline{\lambda_j} \Rightarrow p_i = \overline{p_j}\,. \tag{2.5}$$

As an example, the Fibonacci sequence $0, 1, 1, 2, 3, 5, 8, 13, \ldots$ is a simple linear recurrence sequence of order 2, with recursion rule $f_{n+1} = f_n + f_{n-1}$. It can also be

represented in the form

$$f_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n.$$

For more information on linear recurrence sequences, see [39].

2.5.1 Lower-bounding simple linear recurrence sequences

We are interested in lower-bounding expressions of the form

$$u_n = \sum_{j=1}^{s} \alpha_j \lambda_j^n \tag{2.6}$$

satisfying the property of Equation (2.5), where $\alpha_1, \lambda_1, \ldots, \alpha_s, \lambda_s$ are algebraic integers and $\lambda_1, \ldots, \lambda_s$ have the same absolute value ρ.

Note that eigenvalues of integer-valued matrices are always algebraic integers, and that all algebraic numbers admit an integer multiple that is an algebraic integer. Therefore, assuming the α_j to be algebraic integers will only worsen the bound we derive in this subsection by a constant factor, which will be irrelevant.

In order to make use of Theorem 2.7, it is important to understand the set

$$\{n \in \mathbb{N} : \exists I \subseteq \{1, \ldots, s\}, \sum_{j \in I} \alpha_j \lambda_j^n = 0\} \tag{2.7}$$

The following well-known theorem characterises the set of zeros of linear recurrence sequences. In particular, it gives us a sufficient condition for guaranteeing that the set of zeros of a non-identically zero linear recurrence sequence is finite. Namely, it suffices that the sequence is *non-degenerate*, that is, that no ratio of two of its characteristic roots is a root of unity.

Theorem 2.26 (Skolem-Mahler-Lech)**.** *Let $(u_n)_{n\in\mathbb{N}}$ be a linear recurrence sequence. The set $\{n \in \mathbb{N} : u_n = 0\}$ is a union of a finite set and finitely many arithmetic progressions. Moreover, if u_n is non-degenerate, this set is actually finite.*

It follows from the Skolem-Mahler-Lech theorem that if u_n is non-degenerate then (2.7) must be finite, assuming without loss of generality that

$$\sum_{j \in I} \alpha_j \lambda_j^n$$

is never eventually zero.

We can now apply Theorem 2.7 to get a lower bound on (2.6) that holds for all but finitely many n, by letting K be the splitting field of the characteristic polynomial of u_n, S be the set of prime ideals of the ring of integers of K that appear in the factorisation of each of the algebraic integers α_j and λ_j, and $x_j = \alpha_j \lambda_j^n$ for each j, making (2.6) a sum of S-units.

In the notation of the theorem, we have $Y = \Omega(\rho^n)$. If Λ is an upper bound on the absolute value of the Galois conjugates of each λ_j (that is, each $\sigma_i(\lambda_j)$), then $Z = O(\Lambda^n)$. Thus, for any $\varepsilon > 0$, we know that

$$\sum_{j=1}^{s} \alpha_j \lambda_j^n = \Omega(YZ^{-\varepsilon}) = \Omega\left(\rho^n \Lambda^{-n\varepsilon}\right)$$

Finally, we note that by picking ε to be sufficiently small we can get $\rho\Lambda^{-\varepsilon}$ arbitrarily close to ρ.

Therefore, for any $\eta < \rho$, it is true for sufficiently large n that

$$\left| \sum_{j=1}^{s} \alpha_j \lambda_j^n \right| = \Omega(\eta^n)\,.$$

2.6 Geometry

2.6.1 Convex Polytopes

A *convex polytope* is a subset of $\mathbb{R}^n$ of the form

$$\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} \leq \boldsymbol{b}\},$$

where A is a $d \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^d$. When all the entries of A and coordinates of $\boldsymbol{b}$ are algebraic numbers, the convex polytope $\mathcal{P}$ is said to have an algebraic description.

The decision version of linear programming with canonically-defined algebraic coefficients is in $\exists\mathbb{R}$, as the emptiness of a convex polytope can easily be described by a sentence of the form $\exists x_1 \cdots \exists x_n \sigma(\boldsymbol{x})$.

Finally, we note that even though the decision version of linear programming with rational coefficients is in **PTIME**, allowing algebraic coefficients makes things more complicated. While it has been shown that the decision version of linear programming with canonically-defined algebraic coefficients is solvable in time polynomial in the size of the problem instance and in the degree of the smallest number field containing all algebraic numbers in each instance [1], it turns out that the degree of that extension

can be exponential in the size of the input[2]. In other words, the splitting field of the characteristic polynomial of a matrix can have a degree which is exponential in the degree of the characteristic polynomial, which makes it hard to place this problem in **PTIME**.

We will also make use of the following result:

Theorem 2.27 (Minkowski-Weyl)**.** *Any polytope $\mathcal{P} \subseteq \mathbb{R}^d$ can be written as the sum of two sets $\mathcal{H} \subseteq \mathbb{R}^d$ and $\mathcal{C} \subseteq \mathbb{R}^d$, where $\mathcal{H}$ is a finitely-generated convex hull and $\mathcal{C}$ is a finitely generated cone.*

2.6.2 Fourier-Motzkin Elimination

Fourier-Motzkin elimination is a simple method for solving systems of inequalities. Historically, it was the first algorithm used in solving linear programming, before more efficient procedures such as the simplex algorithm were discovered. The procedure consists in isolating one variable at a time and matching all its lower and upper bounds. Note that this method preserves the set of solutions on the remaining variables, so a solution of the reduced system can always be extended to a solution of the original one.

Theorem 2.28. *It is decidable whether a given convex polytope $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^n : \pi A\boldsymbol{x} < \boldsymbol{b}\}$, where the entries of A are all real algebraic numbers and those of $\boldsymbol{b}$ are real linear forms in logarithms of algebraic numbers, is empty. Moreover, if $\mathcal{P}$ is non-empty one can effectively find a rational vector $\boldsymbol{q} \in \mathcal{P}$.*

Proof. This is done by using Fourier-Motzkin elimination, isolating each term πx_i, instead of just isolating the variable x_i. Note that the coefficients of the terms πx_i will always be algebraic, and the remaining terms will always be linear forms in logarithms of algebraic numbers, which are closed under multiplication by algebraic numbers, and which can be effectively compared by using Theorem 2.11 and Theorem 2.23. □

2.6.3 Integer Points of Convex Semi-algebraic Sets

Due to Theorem 2.22, it is undecidable whether a given semi-algebraic set S in $\mathbb{R}^d$ intersects the integer lattice $\mathbb{Z}^d$. However, if one knows that S is convex, undecidability breaks down, as shown in [51]. This result is central to our approach in Chapter 3.

[2]For example, consider the sequence of field extensions $\mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})$, where p_i denotes the i'th prime.

Theorem 2.29 (Khachiyan and Porkolab)**.** *Let $W \subseteq \mathbb{R}^d$ be a convex semi-algebraic set defined by polynomials of degree at most D and that can be represented in space*[3] *S. In that case, if $W \cap \mathbb{Z}^d \neq \emptyset$, then W must contain an integral point that can be represented in space $SD^{O(d^4)}$.*

2.7 Analysis

2.7.1 Laurent polynomials

A multivariate *Laurent polynomial* is a polynomial in positive and negative powers of variables $z_1, \ldots, z_s$ with complex coefficients. We are interested in Laurent polynomials of the special form

$$g(\boldsymbol{z}) = \sum_{j=1}^{k} \left(c_j {z_1}^{n_{1,j}} \ldots {z_s}^{n_{s,j}} + \overline{c_j} {z_1}^{-n_{1,j}} \ldots {z_s}^{-n_{s,j}} \right)$$

where $c_1, \ldots, c_k \in \mathbb{C}$ and $n_{1,1}, \ldots, n_{s,k} \in \mathbb{Z}$. We call such g *self-conjugate Laurent polynomials*. Notice that if $\boldsymbol{a} \in \mathbb{T}^s$ then $g(\boldsymbol{a})$ is a real number, so we may regard g as a function from $\mathbb{T}^s$ to $\mathbb{R}$.

We say that g is *simple* if g has no constant term and each monomial in g mentions only a single variable.

Lemma 2.30. *Let $g \in \mathbb{C}[z_1^{\pm 1}, \ldots, z_s^{\pm 1}]$ be a self-conjugate Laurent polynomial that has no constant term (not necessarily simple). Given $\boldsymbol{\theta} \in \mathbb{R}^s$ such that $\mathcal{A}(\boldsymbol{\theta}) = \{\mathbf{0}\}$, define a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ by*

$$f(t) = g(\exp(2\pi i t \boldsymbol{\theta}))\,.$$

Then either f is identically zero, or

$$\liminf_{n \to \infty} f(n) < 0\,,$$

where n ranges over the naturals.

Proof. Consider the function $h : \mathbb{R}^s \to \mathbb{R}$ given by $\boldsymbol{x} \mapsto g(\exp(2\pi i \boldsymbol{x}))$. We use an averaging argument to establish that either h is identically zero on $\mathbb{R}^s$ or there exist $\boldsymbol{x}^\star \in [0,1]$ such that $h(\boldsymbol{x}^\star) < 0$.

[3]Here, the space of the representation of a polynomial is defined to be the sum of the base-2 logarithms of its coefficients.

Since, for all non-zero integers n,

$$\int_0^1 \exp(2\pi i n x) dx = 0\,,$$

it follows that

$$\int_{[0,1]^s} h(\boldsymbol{x}) d\boldsymbol{x} = 0\,.$$

Suppose that h is not identically zero over $\mathbb{R}^s$ and hence not identically zero over $[0,1]^s$. Then h cannot be nonnegative on $[0,1]^s$, since the integral over a set of positive measure of a continuous nonnegative function that is not identically zero must be strictly positive. We conclude that there must exist $\boldsymbol{x}^\star \in [0,1]^s$ such that $h(\boldsymbol{x}^\star) < 0$.

By assumption, $\mathcal{A}(\boldsymbol{\theta}) = \{\mathbf{0}\}$. By Corollary 2.25 it follows that

$$\{\exp(n\boldsymbol{\theta}) : n \in \mathbb{N}\}$$

is dense in $\mathbb{T}^s$ and hence has $\exp(2\pi i \boldsymbol{x}^\star)$ as a limit point. Since h is continuous, there are arbitrarily large $n \in \mathbb{N}$ for which

$$f(n) = h(n\boldsymbol{\theta}) \leq \frac{1}{2} h(\boldsymbol{x}^\star) < 0\,,$$

which proves the result. □

Note that this proof could be made constructive by using an effective version of Kronecker's Theorem, as studied in [21] and [59], although we do not make use of this fact in the present thesis.

The following consequence of Lemma 2.30 will be key to proving decidability of the problem at hand. It is a continuous-time extension of [20, Lemma 4].

Theorem 2.31. *Let $g \in \mathbb{C}[z_1^{\pm 1}, \ldots, z_s^{\pm 1}]$ be a simple self-conjugate Laurent polynomial and $\theta_1, \ldots, \theta_s$ be non-zero real numbers. Then either*

$$g(\exp(2\pi i t\theta_1), \ldots, \exp(2\pi i t\theta_s)) = 0 \text{ for all } t \in \mathbb{R}$$

or

$$\liminf_{n\to\infty} g(\exp(2\pi i n\theta_1), \ldots, \exp(2\pi i n\theta_s)) < 0\,,$$

where n ranges over the naturals.

Proof. Note that if $1, \theta_1, \ldots, \theta_s$ are linearly independent over $\mathbb{Q}$ then the result follows from Lemma 2.30. Otherwise, let $\{\theta_{i_1}, \ldots, \theta_{i_k}\}$ be a maximal subset of $\{\theta_1, \ldots, \theta_s\}$ such that $1, \theta_{i_1}, \ldots, \theta_{i_k}$ are linearly independent over $\mathbb{Q}$.

Then, for some $N \in \mathbb{N}$ and each j, one can write

$$N\theta_j = \left(m + \sum_{l=1}^{k} n_l \theta_{i_l} \right) ,$$

where $m, n_1, \ldots, n_k$ are integers that depend on j, whilst N does not depend on j. It follows that for all j and $t \in \mathbb{R}$,

$$\begin{aligned} \exp(2\pi i N\theta_j t) &= \exp(2\pi i m t) \cdot \prod_{l=1}^{k} \exp(2\pi i n_l \theta_{i_l} t) \\ &= \exp(2\pi i t)^m \cdot \prod_{l=1}^{k} \exp(2\pi i \theta_{i_l} t)^{n_l} . \end{aligned}$$

In other words, for all $j \geq k+1$, $\exp(2\pi i N\theta_j t)$ can be written as a product of positive and negative powers of the terms

$$\exp(2\pi i t), \exp(2\pi i \theta_{i_1} t), \ldots, \exp(2\pi i \theta_{i_k} t) .$$

It follows that there exists a self-conjugate Laurent polynomial $h \in \mathbb{C}[z_1^{\pm 1}, \ldots, z_k^{\pm 1}]$, not necessarily simple, but with zero constant term, such that for all $t \in \mathbb{R}$,

$$g(\exp(2\pi i N\theta_1 t), \ldots, \exp(2\pi i N\theta_s t)) = h(\exp(2\pi i t), \exp(2\pi i \theta_{i_1} t), \ldots, \exp(2\pi i \theta_{i_k} t)) .$$

Since $1, \theta_{i_1}, \ldots, \theta_{i_k}$ are linearly independent over $\mathbb{Q}$, the result follows by applying Lemma 2.30 to h. □

In order to compare the asymptotic growth of expressions of the form $t^n \exp(\lambda t)$, for $\lambda \in \mathbb{R}$ and $n \in \mathbb{N}_0$, we define $\prec$ to be the lexicographic order on $\mathbb{R} \times \mathbb{N}_0$, that is,

$$(\eta, j) \prec (\rho, m) \quad \text{iff} \quad \eta < \rho \text{ or } (\eta = \rho \text{ and } j < m) .$$

Clearly $\exp(\eta t) t^j = o(\exp(\rho t) t^m)$ as $t \to \infty$ if and only if $(\eta, j) \prec (\rho, m)$.

Definition 2.1. If $\boldsymbol{b}^T \exp(At)\boldsymbol{v}$ is not identically zero, the maximal $(\rho, m) \in \mathbb{R} \times \mathbb{N}_0$ with respect to $\prec$ for which there is a term $t^m \exp(\lambda t)$ with $\Re(\lambda) = \rho$ in the closed-form expression for $\boldsymbol{b}^T \exp(At)\boldsymbol{v}$ is called *dominant* for $\boldsymbol{b}^T \exp(At)\boldsymbol{v}$.

We now derive a useful corollary of Theorem 2.31:

Corollary 2.32. *Consider a function of the form* $h(t) = \boldsymbol{b}^T \exp(At)\boldsymbol{v}^c$, *where* $A \in \mathbb{R}^{d\times d}$, $\boldsymbol{b}, \boldsymbol{v}^c \in \mathbb{R}^d$, *and* $\boldsymbol{v}^c \in \mathcal{V}^c$, *with* $(\rho, m) \in \mathbb{R} \times \mathbb{N}_0$ *dominant. If* $h(t) \not\equiv 0$, *that is, if it is not identically zero, then we have*

$$-\infty < \liminf_{t\to\infty} \frac{h(t)}{\exp(\rho t)t^m} < 0\,.$$

Proof. Let $\Re(\sigma(A)) = \{\eta \in \mathbb{R} : \eta + i\theta \in \sigma(A), \text{ for some } \theta \in \mathbb{R}\}$. Moreover, for $\eta \in \Re(\sigma(A))$, we define $\boldsymbol{\theta}_\eta = \{\theta \in \mathbb{R}_{>0} : \eta + i\theta \in \sigma(A)\}$. By abuse of notation, we also use $\boldsymbol{\theta}_\eta$ to refer to the vector whose coordinates are exactly the members of this set, ordered in an increasing way. We note that, due to Proposition 2.1 and Proposition 2.2, the following holds:

$$\begin{aligned}
\boldsymbol{b}^T \exp(At)\boldsymbol{v}^c &= \boldsymbol{b}^T \exp(At) \sum_{\eta\in\Re(\sigma(A))} \sum_{\theta\in\boldsymbol{\theta}_\eta} \boldsymbol{v}_{\eta+i\theta} + \boldsymbol{v}_{\eta-i\theta} \\
&= \sum_{\eta\in\Re(\sigma(A))} \sum_{\theta\in\boldsymbol{\theta}_\eta} \boldsymbol{b}^T \exp(At)\boldsymbol{v}_{\eta+i\theta} \\
&\qquad\qquad + \overline{\boldsymbol{b}^T \exp(At)\boldsymbol{v}_{\eta+i\theta}} \\
&= \sum_{\eta\in\Re(\sigma(A))} \sum_{j=0}^{\nu(A)-1} t^j \exp(\eta t) g_{(\eta,j)}(\exp(i\boldsymbol{\theta}_\eta t))
\end{aligned}$$

for some simple self-conjugate Laurent polynomials $g_{(\eta,j)}$. Note that

$$(\rho, m) = \max_{\prec}\{(\eta, j) \in \mathbb{R} \times \mathbb{N}_0 : g_{(\eta,j)}(\exp(i\boldsymbol{\theta}_\eta t)) \not\equiv 0\}\,.$$

The result then follows from Theorem 2.31 and the fact that

$$\liminf_{t\to\infty} \frac{h(t)}{\exp(\rho t)t^m} = \liminf_{t\to\infty} g_{(\rho,m)}(\exp(i\boldsymbol{\theta}_\rho t))\,.$$

□

Chapter 3

Termination of Integer Linear Loops

3.1 Introduction

Termination is a fundamental decision problem in program verification. In particular, termination of programs with linear assignments and linear conditionals has been extensively studied over the last decade. This has led to the development of powerful techniques to prove termination via synthesis of linear ranking functions [14, 19, 27, 36, 73], many of which have been implemented in software verification tools, such as Microsoft's TERMINATOR [37].

A very simple form of imperative programs is *simple linear loops*, that is, programs of the form

$$\mathsf{P1} : \boldsymbol{x} \leftarrow \boldsymbol{u};\ \textit{while}\ B\boldsymbol{x} \geq \boldsymbol{c}\ \textit{do}\ \boldsymbol{x} \leftarrow A\boldsymbol{x} + \boldsymbol{a},$$

where $\boldsymbol{x}$ is vector of variables, $\boldsymbol{u}$, $\boldsymbol{a}$, and $\boldsymbol{c}$ are integer vectors, and A and B are integer matrices of appropriate dimensions. Here the loop guard is a conjunction of linear inequalities and the loop body consists of a simultaneous affine assignment to $\boldsymbol{x}$. If the vectors $\boldsymbol{a}$ and $\boldsymbol{c}$ are both zero then we say that the loop is *homogeneous*.

Suppose that the vector $\boldsymbol{x}$ has dimension d. We say that P1 *terminates* on a set $S \subseteq \mathbb{R}^d$ if it terminates for all initial vectors $\boldsymbol{u} \in S$. Tiwari [88] gave a procedure to decide whether a given simple linear loop terminates on $\mathbb{R}^d$. Later Braverman [20] showed decidability of termination on $\mathbb{Q}^d$. However the most natural problem from the point of view of program verification is termination on $\mathbb{Z}^d$.

While termination on $\mathbb{Z}^d$ reduces to termination on $\mathbb{Q}^d$ in the homogeneous case (by a straightforward scaling argument), termination on $\mathbb{Z}^d$ in the general case is stated as an open problem in [13, 20, 88]. The main result of this chapter is a

procedure to decide termination on $\mathbb{Z}^d$ for simple linear loops when the assignment matrix A is diagonalisable. This represents the first substantial progress on this open problem in over 10 years.

Termination of more complex programs can often be reduced to termination of simple linear loops (see, e.g., [37] or [88, Section 6]). On the other hand, termination becomes undecidable for mild generalisations of simple linear loops, for example, allowing the update function in the loop body to be piecewise linear [13]. An interesting related result is the undecidability of a simple generalisation of the famous Collatz problem [52], as well as the related "FRACTRAN" computational model, introduced by John Horton Conway.

To prove our main result we focus on *eventual non-termination*, where P1 is said to be eventually non-terminating on $\boldsymbol{u} \in \mathbb{Z}^d$ if, starting from initial value $\boldsymbol{u}$, after executing the loop body $\boldsymbol{x} \leftarrow A\boldsymbol{x} + \boldsymbol{a}$ a finite number of times *while disregarding the loop guard* we eventually reach a value on which P1 fails to terminate. Recall that A and $\boldsymbol{a}$ have integer coefficients, so clearly P1 fails to terminate on $\mathbb{Z}^d$ if and only if it is eventually non-terminating on some $\boldsymbol{u} \in \mathbb{Z}^d$.

Given a simple linear loop we show how to compute a convex semi-algebraic set $W \subseteq \mathbb{R}^d$ such that the integer points $\boldsymbol{u} \in W$ are precisely the eventually non-terminating integer initial values. Since it is decidable whether a convex semi-algebraic set contains an integer point [51], we can decide whether an integer linear loop is terminating on $\mathbb{Z}^d$. Convexity is crucial here: deciding whether an arbitrary semi-algebraic set contains any integer points is equivalent to Hilbert's Tenth Problem, and this is undecidable.

Termination over the set of all integer points is easily seen to be co-**NP**-hard. Indeed, if the update function in the loop body is the identity then the loop is non-terminating if and only if there is an integer point satisfying the guard. Thus non-termination subsumes integer programming, which is **NP**-hard.

While our algorithm for deciding termination requires exponential space, it should be noted that the procedure actually solves a more general problem than merely determining the existence of a non-terminating integer point (or, equivalently, the existence of an eventually non-terminating integer point). In fact, up to technicalities, our algorithm computes a representation of the set of all eventually non-terminating integer points. For reference, the closely related problem of deciding termination on the integer points in a given convex polytope is **EXPSPACE**-hard [13]. By contrast, even though not stated explicitly in [88] and [20], deciding termination on $\mathbb{R}^d$ and $\mathbb{Q}^d$ can be reduced in polynomial time to a problem in $\exists\mathbb{R}$, which is within **PSPACE**.

As well as making extensive use of algorithms in real algebraic geometry, the soundness of our decision procedure relies on powerful lower bounds in Diophantine approximation that generalise Roth's Theorem. (The need for such bounds in the inhomogeneous setting was conjectured in the conclusion of [20].) We also use classical results in number theory, such as the Skolem-Mahler-Lech Theorem [54, 58, 85] on linear recurrences. Crucially the well-known and notorious ineffectiveness of Roth's Theorem (and its higher-dimensional and p-adic generalisations) and Skolem-Mahler-Lech Theorem is not a problem for deciding *eventual* non-termination, which is key to our approach.

3.1.1 Related Work

Consider the termination problem for a homogeneous simple linear loop

$$\mathsf{P2} : \boldsymbol{x} \leftarrow \boldsymbol{u};\ \textit{while}\ B\boldsymbol{x} \geq 0\ \textit{do}\ \boldsymbol{x} \leftarrow A\boldsymbol{x}$$

on a single initial value $\boldsymbol{u} \in \mathbb{Z}^d$. Each row $\boldsymbol{b}^T$ of the matrix B corresponds to a loop condition $\boldsymbol{b}^T\boldsymbol{x} \geq 0$. For each such condition, consider the integer sequence $\langle x_n : n \in \mathbb{N}\rangle$ defined by $x_n = \boldsymbol{b}^T A^n \boldsymbol{u}$. Then P2 fails to terminate on an initial value $\boldsymbol{u}$ if and only if each such sequence $\langle x_n\rangle$ is *positive*, i.e., $x_n \geq 0$ for all n. Each sequence $\langle x_n\rangle$ considered above is a *linear recurrence sequence*, as we saw in Section 2.5. Thus deciding whether a homogeneous simple linear loop terminates on a given initial value is at least as hard as the *Positivity Problem* for linear recurrence sequences, that is, the problem of deciding whether a given linear recurrence sequence has exclusively non-negative terms.

The Positivity Problem has been studied at least as far back as the 1970s [10, 43, 56, 79, 80]. Thus far decidability is known only for sequences satisfying recurrences of order 5 or less. It is moreover known that showing decidability at order 6 will necessarily entail breakthroughs in transcendental number theory, specifically significant new results in Diophantine approximation [69]. Even for *simple* linear recurrence sequences, decidability has only been established for order at most 9 [68].

The key difference between studying termination of simple linear loops over $\mathbb{Z}^d$ rather than a single initial value is that the former problem can be approached through eventual termination. In this sense the termination problem is related to the *Ultimate Positivity Problem* for linear recurrence sequences, which asks whether all but finitely many terms of a given sequence are positive [70]. This allows us to bring to bear powerful non-effective Diophantine-approximation techniques, specifically the S-units Theorem of Evertse, van der Poorten, and Schlickewei [40, 89]. Such tools enable us

to obtain decidability of termination for matrices of arbitrary dimension, assuming diagonalisability.

Chonev et al. [31] study higher dimensional versions of Kannan and Lipton's Orbit Problem [50]. These can be seen as versions of the termination problem for linear loops on a fixed initial value. Chonev et al. use substantially different technology from that of this chapter, including Baker's Theorem on linear forms in logarithms [7], and correspondingly relies on restrictions on the dimension of data in problem instances to obtain decidability.

Termination of P1 under the assumption that all eigenvalues of A are real was studied in [76] using spectral techniques. However, as will become clear throughout the course of this chapter, most of the machinery that we use is needed to tackle the case where there are both real and complex eigenvalues with the same absolute value. In the setting of [76], the set of eventually non-terminating points is in fact a polytope, which can be effectively computed resorting only to straightforward linear algebra.

While we use spectral and number-theoretic techniques in this chapter, another well-studied approach for proving termination of linear loops involves designing linear ranking functions, that is, linear functions from the state space to a well-founded domain such that each iteration of the loop strictly decreases the value of the ranking function. However, this approach is incomplete: it is not hard to construct an example of a terminating loop which admits no linear ranking function[1]. Sound and relatively complete methods for synthesising linear ranking functions can be found in [73] and [14]. Whether a linear ranking function exists can be decided in polynomial time when the state space is $\mathbb{Q}^d$ [73] and is co-**NP**-complete when the state space is $\mathbb{Z}^d$ [14].

3.2 Overview of Main Results

The main result of this chapter is as follows:

Theorem 3.1. *The termination over the integers (as defined in Section 3.1) of simple linear loops of the form*

$$\mathsf{P1} : \boldsymbol{x} \leftarrow \boldsymbol{u}\text{; while } B\boldsymbol{x} \geq \boldsymbol{c} \text{ do } \boldsymbol{x} \leftarrow A\boldsymbol{x} + \boldsymbol{a}$$

is decidable in exponential space if A is diagonalisable and in polynomial space if A has dimension at most 4.

[1] For example, consider the program while$(x \geq 0)\{x \leftarrow -2x + 10\}$.

In this section we give a high-level overview of the proof of Theorem 3.1.

Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be the affine function $f(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{a}$ computed by the body of the while loop in P1 and $P = \{\boldsymbol{x} \in \mathbb{R}^d : B\boldsymbol{x} \geq \boldsymbol{c}\}$ the convex polytope corresponding to the loop guard. We define the set of *non-terminating points* to be

$$NT = \{\boldsymbol{u} \in \mathbb{R}^d : \forall n \in \mathbb{N},\ f^n(\boldsymbol{u}) \in P\}\,.$$

Following Braverman [20], we moreover define the set of *eventually non-terminating points* to be

$$ENT = \{\boldsymbol{u} \in \mathbb{R}^d : \exists n \in \mathbb{N},\ f^n(\boldsymbol{u}) \in NT\}\,.$$

It is easily seen from the above definitions that both NT and ENT are convex sets.

By definition, P1 is non-terminating on $\mathbb{Z}^d$ if and only if NT contains an integer point. It is moreover clear that NT contains an integer point if and only if ENT contains an integer point.

Recall that a subset of $\mathbb{R}^d$ is said to be *semi-algebraic* if it is a Boolean combination of sets of the form $\{\boldsymbol{x} \in \mathbb{R}^d : p(\boldsymbol{x}) \geq 0\}$, where p is a polynomial with integer coefficients.

Define $W \subseteq \mathbb{R}^d$ to be a *non-termination witness set* (or simply a witness set) if it satisfies the following two properties (where $\overline{\mathbb{Q}}$ denotes the set of algebraic numbers):

(i) W is convex and semi-algebraic;

(ii) $W \cap \overline{\mathbb{Q}}^d = ENT \cap \overline{\mathbb{Q}}^d$.

The integer points in a witness set W are precisely the integer points of ENT, and so P1 is non-terminating on $\mathbb{Z}^d$ precisely when W contains an integer point. Our approach to solving the termination problem consists in computing a witness set W for a given program and then using Theorem 2.29.

Our approach does not attempt to characterise the set ENT directly, but rather uses the witness set W as a proxy. However, our techniques do allow us to establish that $\mathrm{Cl}(ENT) = \mathrm{Cl}(W)$, which implies that $\mathrm{Cl}(ENT)$ is semi-algebraic, since the closure of a semi-algebraic set is semi-algebraic (see Lemma 2.20). A natural question is whether the set ENT itself is semi-algebraic, which we leave as an open problem.

We next describe some restrictions on linear loops that can be made without loss of generality for deciding termination; that will ease our upcoming analysis.

We first reduce the problem of computing witness sets in the general case to the same problem in the homogeneous case. Note that Program P1 terminates on a given

initial value $\boldsymbol{u} \in \mathbb{Z}^d$ if and only if the homogeneous program P3 below terminates on initial value $(\boldsymbol{u}, 1)$:

$$\mathsf{P3}: \boldsymbol{x} \leftarrow \begin{pmatrix} \boldsymbol{u} \\ 1 \end{pmatrix} \textit{ while } \begin{pmatrix} B & -\boldsymbol{c} \end{pmatrix} \boldsymbol{x} \geq 0 \textit{ do } \boldsymbol{x} \leftarrow \begin{pmatrix} A & \boldsymbol{a} \\ 0 & 1 \end{pmatrix} \boldsymbol{x}$$

Note that if A is diagonalisable then all eigenvalues of $\begin{pmatrix} A & \boldsymbol{a} \\ 0 & 1 \end{pmatrix}$ are simple, with the possible exception of the eigenvalue 1. (Recall that an eigenvalue is said to be simple if it has multiplicity one as a zero of the minimal polynomial of A.) Now if W is a witness set for program P3 then $\left\{ \boldsymbol{u} \in \mathbb{R}^d : \begin{pmatrix} \boldsymbol{u} \\ 1 \end{pmatrix} \in W \right\}$ is a witness set for P1. We conclude that, in order to settle the inhomogeneous case with a diagonalisable matrix, it suffices to compute a witness set in the case of a homogeneous linear loop P2 in which the only repeated eigenvalues of the new matrix A are positive and real. Likewise, to handle the inhomogeneous case for matrices of dimension at most d, it suffices to be able to compute witness sets in the homogeneous case for matrices of dimension at most $d + 1$.[2]

We can further simplify the homogeneous case by restricting to loop guards that consist of a single linear inequality. To see this, first note that program P2 above is eventually non-terminating on $\boldsymbol{u}$ if and only if, for each row $\boldsymbol{b}^T$ of B, program P4 below is eventually non-terminating on $\boldsymbol{u}$:

$$\mathsf{P4}: \boldsymbol{x} \leftarrow \boldsymbol{u};\ \textit{while } \boldsymbol{b}^T \boldsymbol{x} \geq 0 \textit{ do } \boldsymbol{x} \leftarrow A\boldsymbol{x}.$$

Noting that the finite intersection of convex semi-algebraic sets is again convex and semi-algebraic, we can compute a witness set for P2 as the intersection of witness sets for each version of P4.

The final simplification concerns the notion of non-degeneracy. We say that matrix A is *degenerate* if it has distinct eigenvalues $\lambda_1 \neq \lambda_2$ whose quotient λ_1/λ_2 is a root of unity. The *order* of a root of unity ζ is defined as $\min\{n > 0 : \zeta^n = 1\}$.

Given an arbitrary matrix A, let L be the least common multiple of all orders of quotients of distinct eigenvalues of A which are roots of unity. It is known that $L = 2^{O(d\sqrt{\log d})}$ [39, Theorem 1.2]. The looser bound $L = 2^{O(d^3\sqrt{\log(d)})}$, which would suffice for our purposes, can be derived as follows. Let $g(n)$ be the largest least

[2]Note that whilst Braverman [20] shows how to decide termination over the integers for homogeneous programs with arbitrary update matrices, he does *not* compute a witness set for such programs—indeed this remains an open problem since it would enable one to solve termination over the integers for arbitrary inhomogeneous programs.

common multiple of any partition of n; the function $g(n)$ is named after Landau, who showed that

$$\lim_{n\to\infty} \ln(g(n)) = (1+o(1))\sqrt{n\ln(n)}.$$

The number of quotients of eigenvalues of A that are roots of unity is trivially upper bounded by d^2, and the order of each of them is loosely upper bounded by $2d^4$, due to Lemma 2.9 and to the fact that the degree of a quotient of eigenvalues of A is upper bounded by d^2, yielding a $2d^6$ bound on the sums of the orders of all quotients of eigenvalues of A that are roots of unity, which entails the upper bound $L \leq g(2d^6) = 2^{O(d^3\sqrt{\log(d)})}$.

The eigenvalues of the matrix A^L have the form λ^L for λ an eigenvalue of A, by the spectral mapping theorem. It follows that A^L is non-degenerate, since if λ_1, λ_2 are eigenvalues of A such that λ_1^L/λ_2^L is a root of unity then λ_1/λ_2 is a root of unity and hence $\lambda_1^L/\lambda_2^L = 1$. Note that all eigenvectors of A are still eigenvectors of A^L, thus, whenever A is diagonalisable, so will A^L be.

Now program P4 is eventually non-terminating on $\boldsymbol{u} \in \mathbb{Z}^d$ if and only if program P5 below is eventually non-terminating on the set $\{\boldsymbol{u}, A\boldsymbol{u}, \ldots, A^{L-1}\boldsymbol{u}\}$:

$$\mathsf{P5} : \boldsymbol{x} \leftarrow \boldsymbol{v};\ \textit{while}\ \boldsymbol{b}^T\boldsymbol{x} \geq 0\ \textit{do}\ \boldsymbol{x} \leftarrow A^L\boldsymbol{x}.$$

Thus if W is a witness set for P5 then $\bigcap_{i=0}^{L-1}\{\boldsymbol{u} \in \mathbb{Z}^d : A^i\boldsymbol{u} \in W\}$ is a witness set for P4.

The main technical result of the chapter is the following proposition:

Proposition 3.2. *Consider a homogeneous simple linear loop*

$$\mathsf{P4} : \boldsymbol{x} \leftarrow \boldsymbol{u};\ \textit{while}\ \boldsymbol{b}^T\boldsymbol{x} \geq 0\ \textit{do}\ \boldsymbol{x} \leftarrow A\boldsymbol{x},$$

such that A is non-degenerate. If A has dimension at most 5, we can compute a witness set for P4 in polynomial space. In the case where all complex eigenvalues of A are simple, we can compute a witness set for P4 in exponential space.

Bearing in mind that the transformation from P1 to P4 increases the dimension of A by one and does not introduce repeated complex eigenvalues, it follows from Proposition 3.2 that we can also compute witness sets for simple linear loops of the form P1 under the assumptions of Theorem 3.1, and thus we obtain the decidability part of Theorem 3.1. The exponential-space bound in Theorem 3.1 is obtained by bounding the representation of the witness set in Proposition 3.2, as done in Section 3.4.

In the rest of this section we give a brief summary of the proof of Proposition 3.2.

To compute a witness set W for P4 we first partition the eigenvalues of the update matrix A by grouping eigenvalues of equal modulus. Correspondingly we write $\mathbb{R}^d$ as a direct sum $\mathbb{R}^d = V_1 \oplus \cdots \oplus V_m$, where each subspace V_i is the sum of (generalised) eigenspaces of A associated with eigenvalues of the same modulus. Assume that V_1 corresponds to the eigenvalues of maximum modulus, V_2 the next greatest modulus, etc. Then there are two main steps in the construction of W:

1. By analysing multiplicative relationships among eigenvalues of the same modulus, we show that for each subspace V_i the set $ENT \cap V_i$ of eventually non-terminating initial values in V_i is semi-algebraic.

2. Given $\boldsymbol{v} \in \mathbb{R}^d$, we can write $\boldsymbol{v} = \boldsymbol{v}_1 + \cdots + \boldsymbol{v}_m$, with $\boldsymbol{v}_i \in V_i$. Using Theorem 2.7 on S-units, we show that if all entries of $\boldsymbol{v}$ are algebraic numbers then the eventual non-termination of P4 on $\boldsymbol{v}$ is determined by its eventual non-termination on each $\boldsymbol{v}_i$ separately. More precisely we look for the first $\boldsymbol{v}_i$ such that the sequence $\langle \boldsymbol{b}^T A^n \boldsymbol{v}_i : n \in \mathbb{N} \rangle$ is infinitely often non-zero. Then P4 is eventually non-terminating on $\boldsymbol{v}$ if and only if it is eventually non-terminating on $\boldsymbol{v}_i$.

The computability of a witness set W easily follows from items 1 and 2 above. Our techniques require that the update matrix in the original linear loop P1 either be diagonalisable or have dimension at most 4. Eliminating these restrictions seems to require solving the Ultimate Positivity Problem for linear recurrence sequences of order greater than 5, which in turn requires solving hard open problems in the theory of Diophantine approximation [69].

3.3 Algorithm for Universal Termination

Our goal in this section is to prove Proposition 3.2, which we restate below. We have already shown in Section 3.2 that the main result of this chapter, Theorem 3.1, then follows.

Proposition 3.2. *Consider a homogeneous simple linear loop*

$$\mathsf{P4} : \boldsymbol{x} \leftarrow \boldsymbol{u}; \textit{ while } \boldsymbol{b}^T \boldsymbol{x} \geq 0 \textit{ do } \boldsymbol{x} \leftarrow A\boldsymbol{x},$$

such that A is non-degenerate. If A has dimension at most 5, we can compute a witness set for P4 in polynomial space. In the case where all complex eigenvalues of A are simple, we can compute a witness set for P4 in exponential space.

Let A have eigenvalues $\lambda_1, \ldots, \lambda_l$, with respective indices $\nu_1, \ldots, \nu_l$. Given $\boldsymbol{u} \in \mathbb{R}^d$, from our observations in Equation (3.1), we can write

$$\boldsymbol{b}^T A^n \boldsymbol{u} = \sum_{j=1}^{l} \sum_{k=0}^{\nu_j - 1} \boldsymbol{\alpha}_{j,k}^T \boldsymbol{u}\, n^k \lambda_j^n \,, \tag{3.1}$$

where the $\boldsymbol{\alpha}_{j,k}$ are vectors of algebraic numbers that do not depend on $\boldsymbol{u}$, and whose coordinates are all algebraic numbers lying in the extension field of $\mathbb{Q}$ generated by the eigenvalues of A. Moreover, Equation (3.1) holds for all $n \geq d$.

Since the characteristic polynomial of A has integer coefficients (and is monic), the eigenvalues of A are all algebraic integers. Moreover, since for any positive integer $t > 0$ we have that $t \cdot \boldsymbol{b}^T A^n \boldsymbol{u} \geq 0$ if and only if $\boldsymbol{b}^T A^n \boldsymbol{u} \geq 0$, by rescaling we can assume that the vectors $\boldsymbol{\alpha}_{j,k}$ in (3.1) are comprised of algebraic integers.

Now let us partition the eigenvalues of A into sets $S_1, \ldots, S_m$ by grouping eigenvalues of equal modulus. Assume that S_1 contains eigenvalues of maximum modulus, S_2 eigenvalues of the next greatest modulus, etc. Correspondingly we write $\mathbb{R}^d$ as a direct sum of subspaces $\mathbb{R}^d = V_1 \oplus \cdots \oplus V_m$, where each subspace V_i is the sum of (generalised) eigenspaces of A associated with eigenvalues in S_i. By the assumption that A is non-degenerate, i.e., that no quotient of two distinct eigenvalues is a root of unity, S_i cannot have both a positive and a negative real eigenvalue of the same modulus. Thus each set S_i contains at most one real eigenvalue.

3.3.1 Eventual Non-Termination on Subspace V_i

We first consider the eventual non-termination of P4 on initial vectors in the subspace V_i for a fixed $i \in \{1, \ldots, m\}$. Writing

$$ENT_i := ENT \cap V_i,$$

our goal is to show that ENT_i is semi-algebraic.

Given $\boldsymbol{u} \in V_i$, membership of $\boldsymbol{u}$ in ENT_i can be characterised in terms of the *Ultimate Positivity* of the sequence $\langle\, \boldsymbol{b}^T A^n \boldsymbol{u} : n \in \mathbb{N} \,\rangle$. More precisely, $\boldsymbol{u} \in ENT_i$ if and only if $\boldsymbol{b}^T A^n \boldsymbol{u} \geq 0$ for all but finitely many n. In particular, defining

$$ZERO := \{\boldsymbol{u} \in \mathbb{R}^d : \forall n \geq d,\, \boldsymbol{b}^T A^n \boldsymbol{u} = 0\}$$

and $ZERO_i := ZERO \cap V_i$, we have that $ZERO_i \subseteq ENT_i$.

It is easy to see that $ZERO_i$ is semi-algebraic. Indeed the uniqueness part of [44, Proposition 2.11] implies that $\boldsymbol{b}^T A^n \boldsymbol{u} = 0$ for all $n \geq d$ if and only if each term $n^k \lambda_j^n$ has coefficient zero in the expression (3.1). Thus

$$ZERO = \left\{ \boldsymbol{u} \in \mathbb{R}^d : \bigwedge_{j=1}^{l} \bigwedge_{k=0}^{\nu_j - 1} \boldsymbol{\alpha}_{j,k}^T \boldsymbol{u} = 0 \right\} .$$

is semi-algebraic. Since V_i is a semi-algebraic subset of $\mathbb{R}^d$, being spanned by a subset of the columns of P, it follows that $ZERO_i$ is semi-algebraic.

Proposition 3.3. *For each $i \in \{1, \ldots, m\}$, the set ENT_i is semi-algebraic.*

Proof. We consider three (overlapping) cases. Under the hypotheses of Proposition 3.2 at least one of these cases will apply.

Case I: A has dimension at most 5. Assume that A has dimension at most 5. The situations in which S_i does not contain a positive real eigenvalue, or all of the complex eigenvalues in S_i are simple, will be handled under Cases II and III, below. Otherwise, let $\lambda \in S_i$ be a complex eigenvalue of index at least 2. Since A has dimension at most 5, it must be the case that λ and its complex conjugate $\overline{\lambda}$ both have index exactly 2. Let $\rho \in S_i$ be the positive real eigenvalue. Since A has dimension at most 5, ρ must be simple. Thus $S_i = \{\rho, \lambda, \overline{\lambda}\}$ contains all the eigenvalues of A.

For $\boldsymbol{u} \in V_i$ we can write

$$\boldsymbol{b}^T A^n \boldsymbol{u} = \left(\boldsymbol{\alpha}_0 \rho^n + (\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 n)\lambda^n + \overline{(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 n)\lambda^n} \right)^T \boldsymbol{u} ,$$

for all $n \geq d$, where $\boldsymbol{\alpha}_0$ is a vector of real algebraic numbers, $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$ are vectors of complex algebraic numbers.

If $\boldsymbol{\beta}_1^T \boldsymbol{u} \neq 0$, then as n tends to infinity the dominant terms on the right-hand side above are constant multiples of $n\lambda^n$ and $n\overline{\lambda^n}$. In this case it follows from Theorem 2.31 that $\boldsymbol{b}^T A^n \boldsymbol{u}$ gets arbitrarily large negative values as n grows, and hence $\boldsymbol{u} \notin ENT_i$.

In case $\boldsymbol{\beta}_1^T \boldsymbol{u} = 0$, the argument is a simpler version of the approach in Case III; however, we include the details of this special case, since the reader may find them instructive.

Define $f : \mathbb{T} \to \mathbb{R}$ by

$$f(z) = \boldsymbol{\alpha}_0^T \boldsymbol{u} + \boldsymbol{\beta}_0^T \boldsymbol{u} z + \overline{\boldsymbol{\beta}_0^T \boldsymbol{u} z} .$$

Then $\boldsymbol{b}^T A^n \boldsymbol{u} = \rho^n f(\lambda^n / \rho^n)$ for all $n \geq d$.

Since A is assumed to be non-degenerate, λ/ρ is not a root of unity. Thus $\{\lambda^n/\rho^n : n \in \mathbb{N}\}$ is dense in $\mathbb{T}$. It follows that $\boldsymbol{u} \in ENT$ if and only if $f(z) \geq 0$ for all $z \in \mathbb{T}$. By inspection this last condition is equivalent to $\boldsymbol{\alpha}_0^T \boldsymbol{u} \geq 2|\boldsymbol{\beta}_0^T \boldsymbol{u}|$. We conclude that

$$ENT_i = \left\{ \boldsymbol{u} \in V_i : \boldsymbol{\beta}_1^T \boldsymbol{u} = 0 \wedge \boldsymbol{\alpha}_0^T \boldsymbol{u} \geq 2|\boldsymbol{\beta}_0^T \boldsymbol{u}| \right\} ,$$

and hence ENT_i is semi-algebraic.

Case II: S_i does not contain a positive real eigenvalue. It follows from Theorem 2.31 that, if S_i does not contain a positive real eigenvalue, then for $\boldsymbol{u} \in V_i$ the sequence $\boldsymbol{b}^T A^n \boldsymbol{u}$ is either identically zero for $n \geq d$ or is infinitely often strictly positive and infinitely often strictly negative. Thus in this case $ENT_i = ZERO_i$. But we have already shown that $ZERO_i$ is semi-algebraic.

Case III: all complex eigenvalues in S_i are simple. Suppose that all complex eigenvalues in S_i are simple. If S_i contains no positive real eigenvalue then Case II applies. Thus we may assume that S_i comprises a positive real eigenvalue ρ of index t and simple complex eigenvalues $\lambda_1, \overline{\lambda_1}, \ldots, \lambda_s, \overline{\lambda_s}$. Given $\boldsymbol{u} \in V_i$ we can write

$$\begin{aligned} \boldsymbol{b}^T A^n \boldsymbol{u} &= \boldsymbol{b}^T P^{-1} J^n P \boldsymbol{u} \\ &= \left[\sum_{j=0}^{t-1} \boldsymbol{\alpha}_j n^j \rho^n + \sum_{j=1}^{s} (\boldsymbol{\beta}_j \lambda_j^n + \overline{\boldsymbol{\beta}_j \lambda_j^n}) \right]^T \boldsymbol{u} , \end{aligned} \tag{3.2}$$

where the $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are d-dimensional vectors of algebraic numbers, with all coefficients of each $\boldsymbol{\alpha}_j$ being real.

Since $\rho = |\lambda_1| = \cdots = |\lambda_s|$, if $\boldsymbol{\alpha}_j^T \boldsymbol{u} \neq 0$ for some strictly positive index j, then, for the largest such index j, the term $n^j \rho^n \boldsymbol{\alpha}_j^T \boldsymbol{u}$ is dominating on the right-hand side of (3.2). In particular, if $\boldsymbol{\alpha}_j^T \boldsymbol{u} > 0$ then the sequence $\boldsymbol{b}^T A^n \boldsymbol{u}$ is ultimately positive (hence $\boldsymbol{u} \in ENT_i$), and if $\boldsymbol{\alpha}_j^T \boldsymbol{u} < 0$ then $\boldsymbol{b}^T A^n \boldsymbol{u}$ is not ultimately positive (hence $\boldsymbol{u} \notin ENT_i$). It follows that

$$\left\{ \boldsymbol{u} \in V_i : \bigvee_{j=1}^{t-1} \bigwedge_{k=j+1}^{t-1} (\boldsymbol{\alpha}_j^T \boldsymbol{u} > 0 \wedge \boldsymbol{\alpha}_k^T \boldsymbol{u} = 0) \right\} \tag{3.3}$$

is a subset of ENT_i.

The case that $\boldsymbol{\alpha}_j^T \boldsymbol{u} = 0$ for all $j = 1, \ldots, t-1$ is more subtle since there is no single dominant term in (3.2); this is where we employ the results of Section 2.4 on multiplicative relations. In this case we rewrite (3.2) as

$$\boldsymbol{b}^T A^n \boldsymbol{u} = \rho^n f\left(\frac{\lambda_1^n}{\rho^n}, \ldots, \frac{\lambda_s^n}{\rho^n} \right)^T \boldsymbol{u} , \tag{3.4}$$

where $f : \mathbb{T}^s \to \mathbb{R}^d$ is defined by

$$f(z_1, \ldots, z_s) = \boldsymbol{\alpha}_0 + \sum_{j=1}^{s} \boldsymbol{\beta}_j z_j + \overline{\boldsymbol{\beta}_j z_j}\,.$$

Defining $\boldsymbol{\mu} = (\lambda_1/\rho, \ldots, \lambda_s/\rho)$, we furthermore rewrite (3.4) as

$$\boldsymbol{b}^T A^n \boldsymbol{u} = \rho^n f(\boldsymbol{\mu}^n)^T \boldsymbol{u}\,. \tag{3.5}$$

By Theorem 2.24, $\{\boldsymbol{\mu}^n : n \in \mathbb{N}\}$ is a dense subset of the torus $T(\boldsymbol{\mu})$. Thus the right-hand side of (3.5) is non-negative for every n if and only if $f(\boldsymbol{z})^T \boldsymbol{u} \geq 0$ for all $\boldsymbol{z} \in T(\boldsymbol{\mu})$. It follows that

$$\left\{\boldsymbol{u} \in V_i : \forall \boldsymbol{z} \in T(\boldsymbol{\mu}),\, f(\boldsymbol{z})^T \boldsymbol{u} \geq 0\right\}. \tag{3.6}$$

is a subset of ENT_i.

In Section 2.4 we observed that the set $T(\boldsymbol{\mu})$ was (effectively) semi-algebraic. It follows that we can express the condition $\forall \boldsymbol{z} \in T(\boldsymbol{\mu})$, $f(\boldsymbol{z})^T \boldsymbol{u} \geq 0$ in the first-order theory of the reals. By Theorem 2.16, the set of $\boldsymbol{u} \in \mathbb{R}^d$ satisfying this condition is semi-algebraic. But now ENT_i is the union of the two semi-algebraic sets (3.3) and (3.6), and therefore ENT_i is itself semi-algebraic. □

3.3.2 Definition of a Witness Set

Having shown that $ZERO_i$ and ENT_i are semi-algebraic sets for $i = 1, \ldots, m$, we now define a witness set W for the loop P4.

Given $\boldsymbol{u} \in \mathbb{R}^d$, write $\boldsymbol{u} = \boldsymbol{u}_1 + \cdots + \boldsymbol{u}_m$, with $\boldsymbol{u}_1 \in V_1, \ldots, \boldsymbol{u}_m \in V_m$. Say that $\boldsymbol{u}_i$ is the *dominant component* of $\boldsymbol{u}$ if $\boldsymbol{u}_i \notin ZERO_i$ and $\boldsymbol{u}_j \in ZERO_j$ for all $j < i$. The intuition is that if $\boldsymbol{u}_i$ is dominant then the eventual non-termination of P4 on $\boldsymbol{u}$ is determined by its eventual non-termination on $\boldsymbol{u}_i$. However, to prove this we need to assume $\boldsymbol{u} \in (\overline{\mathbb{Q}} \cap \mathbb{R})^d$. Formally, we have

Proposition 3.4. *If $\boldsymbol{u}_i$ is the dominant component of $\boldsymbol{u} \in (\overline{\mathbb{Q}} \cap \mathbb{R})^d$ then $\boldsymbol{u} \in ENT$ if and only if $\boldsymbol{u}_i \in ENT$.*

Proof. From the fact that $\boldsymbol{u}_i$ is dominant, we have

$$\begin{aligned} \boldsymbol{b}^T A^n \boldsymbol{u} &= \boldsymbol{b}^T A^n (\boldsymbol{u}_1 + \cdots + \boldsymbol{u}_m) \\ &= \boldsymbol{b}^T A^n (\boldsymbol{u}_i + \cdots + \boldsymbol{u}_m) \end{aligned} \tag{3.7}$$

for all $n \geq d$. Moreover, for each $j > i$ it is clear that $|\boldsymbol{b}^T A^n \boldsymbol{u}_j| = O(n^d \rho_j^n)$, where $\rho_j \geq 0$ is the modulus of the eigenvalues in S_j.

We now consider three cases, mirroring the proof of Proposition 3.3.

The first case is that A has dimension at most 5. As observed in the proof of Proposition 3.3, all instances of this case that are not already covered by the second and third cases are such that S_i contains all the eigenvalues of A, and hence $\boldsymbol{u}_i = \boldsymbol{u}$. In this situation the proposition holds trivially.

The second case is that S_i does not contain a positive real eigenvalue. Then it follows from Theorem 2.31 that there is a constant $c < 0$ such that $\boldsymbol{b}^T A^n \boldsymbol{u}_i < c\rho_i^n$ for infinitely many n. In this case neither $\boldsymbol{u}_i$ nor $\boldsymbol{u}$ are elements of ENT.

It remains to consider the case that all complex eigenvalues in S_i are simple. Suppose that the dominant term in the expression for $\boldsymbol{b}^T A^n \boldsymbol{u}_i$ has the form $\alpha n^k \rho_i^n$ for some real constant $\alpha \neq 0$ and $k > 0$. If $\alpha > 0$ then both $\boldsymbol{u}$ and $\boldsymbol{u}_i$ are in ENT and if $\alpha < 0$ then neither $\boldsymbol{u}$ or $\boldsymbol{u}_i$ are in ENT.

Otherwise, the positive real eigenvalue in S_i is also simple. Specialising the expression (3.1) to the case at hand, we have that

$$\boldsymbol{b}^T A^n \boldsymbol{u}_i = \alpha_0 \rho_i^n + \sum_{j=1}^{s} \beta_j \lambda_j^n + \overline{\beta_j \lambda_j^n} \tag{3.8}$$

where α_0 and the β_j are algebraic-integer constants and $\rho_i, \lambda_1, \overline{\lambda_1}, \ldots, \lambda_s, \overline{\lambda_s} \in S_i$. In this case, letting $\eta = (\rho_i + \rho_{i+1})/2$ (note that $\rho_{i+1} < \eta < \rho_i$), one can use Theorem 2.7 as we did in Section 2.5.1 to show that

$$\boldsymbol{b}^T A^n \boldsymbol{u}_i = \Omega(\eta^n).$$

Since it holds trivially that, for all $j > i$, $\boldsymbol{b}^T A^n \boldsymbol{u}_j = o(\eta^n)$, it follows that

$$|\boldsymbol{b}^T A^n \boldsymbol{u}_j| = o\left(|\boldsymbol{b}^T A^n \boldsymbol{u}_i|\right).$$

Therefore, $\boldsymbol{u} \in ENT$ if and only if $\boldsymbol{u}_i \in ENT$. □

Now we define a witness set W for program P4 by

$$W := \bigcup_{i=1}^{m} \{\boldsymbol{u} \in \mathbb{R}^d : \boldsymbol{u}_i \text{ is the dominant component of } \boldsymbol{u}, \\ \boldsymbol{u}_i \in ENT\} \cup ZERO\,.$$

From the fact that $ZERO_i$, ENT_i, and V_i are semi-algebraic for $i = 1, \ldots, m$, it is easy to see that W is semi-algebraic. It moreover follows from Proposition 3.4 that $W \cap \overline{\mathbb{Q}}^d = ENT \cap \overline{\mathbb{Q}}^d$.

To conclude the proof of Proposition 3.2, it remains to observe that the witness set W, like the actual set ENT of eventually non-terminating points, is convex.

Proposition 3.5. *The witness set W is convex.*

Proof. Suppose $\boldsymbol{y}, \boldsymbol{z} \in W$ and let $\boldsymbol{x} = \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{z}$, where $0 < \lambda < 1$. Moreover, write $\boldsymbol{x} = \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_m$, where $\boldsymbol{x}_1 \in V_1, \ldots, \boldsymbol{x}_m \in V_m$, and likewise for $\boldsymbol{y}$ and $\boldsymbol{z}$.

If $\boldsymbol{y}, \boldsymbol{z} \in \mathit{ZERO}$ then $\boldsymbol{x} \in \mathit{ZERO}$ since the latter is a convex set.

Suppose that $\boldsymbol{y} \in \mathit{ZERO}$ and $\boldsymbol{z}_i \in \mathit{ENT}$ is dominant for $\boldsymbol{z}$ for some index $i \in \{1, \ldots, m\}$. Then $\boldsymbol{x}_i$ is dominant for $\boldsymbol{x}$, and $\boldsymbol{x}_i \in \mathit{ENT}$. Thus $\boldsymbol{x} \in W$.

Otherwise, let $\boldsymbol{y}_i$ be dominant for $\boldsymbol{y}$ and $\boldsymbol{z}_j$ be dominant for $\boldsymbol{z}$ for some $i, j \in \{1, \ldots, m\}$. Then $\boldsymbol{x}_k \in \mathit{ZERO}_k$ for all $k < \min\{i, j\}$ since ZERO_k is convex. Moreover if $k = \min\{i, j\}$ then $\boldsymbol{y}_k, \boldsymbol{z}_k \in \mathit{ENT}_k$, and hence $\boldsymbol{x}_k \in \mathit{ENT}_k$ by convexity of ENT_k. It follows that $\boldsymbol{x} \in W$. □

This concludes the proof of Proposition 3.2. In the remaining part of this section we show that the topological closures of ENT and W are equal.

The inclusion $\mathrm{Cl}(W) \subseteq \mathrm{Cl}(\mathit{ENT})$ can be shown using the fact that the set of algebraic points in any semi-algebraic set is dense in that set, as shown in Lemma 2.19. From this, we get

$$\mathrm{Cl}(W) = \mathrm{Cl}(W \cap \overline{\mathbb{Q}}^d) = \mathrm{Cl}(\mathit{ENT} \cap \overline{\mathbb{Q}}^d) \subseteq \mathrm{Cl}(\mathit{ENT}) \cap \mathrm{Cl}(\overline{\mathbb{Q}}^d) = \mathrm{Cl}(\mathit{ENT}).$$

The reverse inclusion, $\mathrm{Cl}(\mathit{ENT}) \subseteq \mathrm{Cl}(W)$, can be shown in a similar fashion, but this time using the fact that $\mathit{ENT} \cap \overline{\mathbb{Q}}^d$ is dense in ENT. Our remaining goal is this last fact, which is established in Corollary 3.7 below.

We have previously shown that a vector of algebraic numbers $\boldsymbol{u} \in (\overline{\mathbb{Q}} \cap \mathbb{R})^d$ is eventually non-terminating if and only if its dominant component $\boldsymbol{u}_i$ is eventually non-terminating. We now prove a partial result of this nature for general vectors $\boldsymbol{u} \in \mathbb{R}^d$ (that is, dropping the algebraicity assumption).

Proposition 3.6. *Suppose that $\boldsymbol{u} = \boldsymbol{u}_1 + \cdots + \boldsymbol{u}_m \in \mathbb{R}^d$, where $\boldsymbol{u}_1 \in V_1, \ldots, \boldsymbol{u}_m \in V_m$. Then $\boldsymbol{u} \in \mathit{ENT}$ implies that its dominant component $\boldsymbol{u}_i$ is also in ENT.*

Proof. The only non-trivial case corresponds to the situation in which $\boldsymbol{b}^T A^n \boldsymbol{u}_i$ is of the form (3.8). Let f and $\boldsymbol{\mu}$ be as in (3.4), that is, so that $\boldsymbol{b}^T A^n \boldsymbol{u}_i = \rho_i^n f(\boldsymbol{\mu}^n)^T \boldsymbol{u}_i$. If $\boldsymbol{u}_i \notin \mathit{ENT}$, then there exists some constant $c < 0$ and some $\boldsymbol{z} \in T(\boldsymbol{\mu})$ such that $f(\boldsymbol{z})^T \boldsymbol{u}_i = c$. Therefore, for any $\varepsilon > 0$, $\boldsymbol{b}^T A^n \boldsymbol{u}_i < (c + \varepsilon)\rho_i^n$ holds for infinitely many n, due to Theorem 2.24 and to continuity of f, and so $\boldsymbol{u} \notin \mathit{ENT}$. □

Corollary 3.7. *$\mathit{ENT} \cap \overline{\mathbb{Q}}^d$ is dense in ENT.*

Proof. At several points we will rely on the fact that if $X \subseteq \mathbb{R}^d$ is semi-algebraic, then the algebraic points in X are dense in X, again due to Lemma 2.19.

Fix $\boldsymbol{u} \in \mathit{ENT}$ and let $\varepsilon > 0$ be given. We will find $\boldsymbol{v} \in \mathit{ENT} \cap \overline{\mathbb{Q}}^d$ such that $||\boldsymbol{u} - \boldsymbol{v}|| < \varepsilon$.

The case in which $\boldsymbol{u} \in \mathit{ZERO}$ is easy since ZERO is semi-algebraic and so we can take $\boldsymbol{v}$ to be an algebraic point in ZERO that is suitably close to $\boldsymbol{u}$.

Suppose now that $\boldsymbol{u} = \boldsymbol{u}_1 + \cdots + \boldsymbol{u}_m$, where $\boldsymbol{u}_1 \in V_1, \ldots, \boldsymbol{u}_m \in V_m$, with $\boldsymbol{u}_i$ the dominant component of $\boldsymbol{u}$. By Proposition 3.6, $\boldsymbol{u} \in \mathit{ENT}$ implies that $\boldsymbol{u}_i \in \mathit{ENT}$. Since $\mathit{ENT} \cap V_i$ is semi-algebraic, we can pick $\boldsymbol{v}_i \in \mathit{ENT}_i \cap \overline{\mathbb{Q}}^d$ such that

$$\|\boldsymbol{v_i} - \boldsymbol{u_i}\| < \frac{\varepsilon}{n}.$$

Similarly, for each $j > i$, we pick some $\boldsymbol{v}_j \in V_j \cap \overline{\mathbb{Q}}^d$ for which

$$\|\boldsymbol{v_j} - \boldsymbol{u_j}\| < \frac{\varepsilon}{n},$$

and for each $j < i$ we pick some $v_j \in \mathit{ZERO}_j \cap \overline{\mathbb{Q}}^d$ for which

$$\|\boldsymbol{v_j} - \boldsymbol{u_j}\| < \frac{\varepsilon}{n}.$$

Then, letting $\boldsymbol{v} = \boldsymbol{v}_1 + \cdots + \boldsymbol{v}_m \in \overline{\mathbb{Q}}^d$, it follows that $\|\boldsymbol{u} - \boldsymbol{v}\| < \varepsilon$. Finally, by Proposition 3.4 we have $\boldsymbol{v} \in \mathit{ENT}$ since $\boldsymbol{v}_i$ is the dominant component of $\boldsymbol{v}$ and $\boldsymbol{v}_i \in \mathit{ENT}$ by construction. □

3.4 Complexity Analysis

The purpose of this section is to justify our previous claims about the complexity of the algorithm presented in this chapter. We do this by proving the following result.

Proposition 3.8. *Our procedure requires space* $\mathit{poly}(\log \max_{i,j}|A_{ij}|, d)^{\mathit{poly}(d)}$.

Proof. There are three critical steps in our procedure for which a super-polynomial amount of space is required: when reducing to the case in which A is non-degenerate, when performing quantifier elimination, and when testing whether the witness set W intersects the integer lattice.

The last of these steps runs in space $SD^{O(d^4)}$, where S denotes the size of the representation of the quantifier-free formula defining the witness set W, D denotes the maximum degree of the polynomials occurring in that formula, and d denotes the dimension of the ambient space. Since d remains fixed throughout the procedure (apart

from an increase by 1 in the reduction to the homogeneous case), it remains to show that S and D are bounded by an expression of the form $poly(\log \max_{i,j}|A_{ij}|, d)^{poly(d)}$.

The reduction to the case in which A is non-degenerate entails an increase by a factor of $poly(\log \max_{i,j}|A_{ij}|, d)^{poly(d)}$ in the size of the formula defining the witness set W, as the least common multiple of the orders of all ratios of eigenvalues of A that are roots of unity is $L = 2^{O(d\sqrt{\log d})}$ and $\log \max_{i,j}|A_{ij}^L| \leq \log(d^L \max_{i,j}|A_{ij}|^L) = L \log(d \max_{i,j}|A_{ij}|)$.

It remains to show that the quantifier-free formula defining the witness set W in the case where A is non-degenerate takes space $poly(\log \max_{i,j}|A_{ij}|, d)^{poly(d)}$ and involves exclusively polynomials of degree $poly(\log \max_{i,j}|A_{ij}|, d)^{poly(d)}$.

Let D_0, H_0 denote the maximum degree and height across all the eigenvalues of A, respectively. Then $D_0 \leq d$ and $\log H_0 \leq \log(d! \max_{i,j}|A_{ij}|^d) \leq d \log(d \max_{i,j}|A_{ij}|)$. Before performing quantifier elimination, the degree of any polynomial in the defining formula of the witness set W is bounded by $(D_0 \log H_0)^{O(d^2)}$, and the number of such polynomials is bounded by $O(d)$, by Masser's theorem. Finally, after applying quantifier elimination, we know that $D \leq (D_0 \log H_0)^{O(d^3)}$ and that $S \leq d^{O(d^2)}(D_0 \log H_0)^{O(d^4)}$, thanks to Theorem 2.17. □

3.5 Conclusion

We have shown decidability of termination of simple linear loops over the integers under the assumption that the update matrix is diagonalisable, partially answering an open problem of [88, 20]. As we have explained before, the termination problem on the same class of linear loops, but for fixed initial values, seems to have a different character and to be more difficult. In this respect it is interesting to note that there are other settings in which universal termination is an easier problem than pointwise termination. For example, universal termination of Petri nets (also known as *structural boundedness*) is **PTIME**-decidable, but the pointwise termination problem is **EXPSPACE**-hard.

Braverman conjectured in [20] that it would be necessary that one be able to decide pointwise termination in order to decide universal termination over the integers. By contrast, the approach we have taken in this chapter has focused *eventual* termination: even for *simple* linear loops, the question of pointwise termination remains open.

A natural subject for further work would be the extension of our result to instances with non-diagonalisable matrices, or showing that there are unavoidable number-theoretic obstacles to proving decidability of this problem, as is the case with Ultimate

Positivity [69]. Another relevant problem would be the computational complexity of the termination problem. While there is a large gap between the co-**NP** lower complexity bound mentioned in Section 3.1 and the **EXPSPACE** upper bound of our procedure, this may be connected with the fact that our procedure computes a representation of the set of all integer eventually non-terminating points.

Finally, there is the question of whether the respective sets of terminating and non-terminating points are semi-algebraic. Note that an *effective* semi-algebraic characterisation of the set of terminating points would allow us to solve the termination problem over fixed initial values.

Chapter 4

Solvability of Matrix-Exponential Equations

4.1 Introduction

Reachability problems are a staple of theoretical computer science and verification, one of the best-known examples being the Halting Problem for Turing machines. In this chapter, our motivation originates from systems that evolve continuously subject to linear differential equations; such objects arise in the analysis of a range of models, including linear hybrid automata, continuous-time Markov chains, linear dynamical systems and cyber-physical systems as they are used in the physical sciences and engineering—see, e.g., [3].

More precisely, consider a system consisting of a finite number of discrete locations (or control states), having the property that the continuous variables of interest evolve in each location according to some linear differential equation of the form $\dot{\boldsymbol{x}} = A\boldsymbol{x}$; here $\boldsymbol{x}$ is a vector of continuous variables, and A is a square 'rate' matrix of appropriate dimension.[1] As is well-known, in each location the closed form solution $\boldsymbol{x}(t)$ to the differential equation admits a matrix-exponential representation of the form $\boldsymbol{x}(t) = \exp(At)\boldsymbol{x}(0)$. Thus if a system evolves through a series of k locations, each with rate matrix A_i, and spending time $t_i \geq 0$ in each location, the overall effect on the initial configuration $\boldsymbol{x}(0)$ is determined by multiplying by the matrix

$$\prod_{i=1}^{k} \exp(A_i t_i)\,.$$

A particularly interesting situation arises when the matrices A_i commute; in such cases, one can show that the order in which the locations are visited (or indeed

[1] In this motivating example, we assume that there are no discrete resets of the continuous variables when transitioning between locations.

whether they are visited only once or several times) is immaterial, the only relevant data being the total time spent in each location. Natural questions then arise as to what kinds of linear transformations can thus be achieved by such systems.

The main object of study of this chapter will be the following decision problem:

Definition 4.1. Given square matrices $A_1, \ldots, A_k$ and C, all of the same dimension, whose entries are real algebraic numbers, the *Matrix-Exponential Semigroup Problem* (MESP) asks whether C a is member of the matrix semigroup generated by

$$\{\exp(A_i t_i) : t_i \in \mathbb{R}, t_i \geq 0, i = 1, \ldots, k\}.$$

We will show that MESP is undecidable in general, but decidable when the matrices $A_1, \ldots, A_k$ commute.

4.1.1 Related Work

In this section, we describe a few decision problems that can be seen as discrete-time analogues of the questions we deal with in this chapter.

Given $k+1$ square matrices $A_1, \ldots, A_k, C$, all of the same dimension, and with algebraic entries, the *Matrix Semigroup Membership Problem* consists in deciding whether the matrix C belongs to the multiplicative semigroup generated by $A_1, \ldots, A_k$. On the other hand, given the multiplicative matrix equation

$$\prod_{i=1}^{k} A_i^{n_i} = C\,,$$

one may be interested in determining whether it admits a solution $n_1, \ldots, n_k \in \mathbb{N}$.

In general, both problems have been shown to be undecidable, in [72] and [11]. When the matrices $A_1, \ldots, A_k$ commute, the two problems are equivalent, and known to be decidable [4].

Prior to [4], the case $k = 1$ was shown to be decidable in [50] and the case $k = 2$ was shown to be decidable in [23].

See [42] for a relevant survey, and [29] for some interesting related problems.

The following continuous analogue of Kannan and Lipton's Orbit Problem [50] was shown to be decidable in [41]:

Definition 4.2 (Continuous Orbit Problem)**.** Given an $n \times n$ matrix A with algebraic entries and two n-dimensional vectors $\boldsymbol{x}, \boldsymbol{y}$ with algebraic coordinates, does there exist a non-negative real number t such that $\exp(At)\boldsymbol{x} = \boldsymbol{y}$?

The paper [28] simplifies the argument of [41] and shows polynomial-time decidability. Moreover, a continuous version of Skolem's Problem was dealt with in [12], where a decidability result is presented for some instances of the problem.

As mentioned earlier, an important motivation for our work comes from the analysis of hybrid automata. In addition to [3], excellent background references on the topic are [49, 48].

4.1.2 Decision Problems

We start by defining three decision problems that will be central to this chapter: the *Matrix-Exponential Problem*, the *Linear-Exponential Problem*, and the *Algebraic-Logarithmic Integer Programming* problem. The relation between these problems will be outlined in Section 4.1.3, but the reader should keep in mind that our ultimate objective is the study of the *Matrix-Exponential Semigroup Problem*, stated in Definition 4.1.

Definition 4.3. Given square matrices $A_1, \ldots, A_k$ and C, all of the same dimension and with real algebraic entries, the *Matrix-Exponential Problem* (MEP) consists in determining whether there exist real numbers $t_1, \ldots, t_k \geq 0$ such that

$$\prod_{i=1}^{k} \exp(A_i t_i) = C\,. \tag{4.1}$$

We will also consider a generalised version of this problem, called the *Generalised MEP*, in which the matrices $A_1, \ldots, A_k$ and C are allowed to have complex algebraic entries and in which the input to the problem also mentions a polytope $\mathcal{P} \subseteq \mathbb{R}^{2k}$ that is specified by linear inequalities with real algebraic coefficients. In the generalised problem we seek $t_1, \ldots, t_k \in \mathbb{C}$ that satisfy (4.1) and such that the vector $(\Re(t_1), \ldots, \Re(t_k), \Im(t_1), \ldots, \Im(t_k))$ lies in $\mathcal{P}$.

In the case of commuting matrices, the Generalised Matrix-Exponential Problem can be analysed block-wise, as we shall see in Section 4.3, leading us to the following problem:

Definition 4.4. An instance of the *Linear-Exponential Problem* (LEP) consists of a system of equations

$$\exp\left(\sum_{i \in I} \lambda_i^{(j)} t_i\right) = c_j \exp(d_j) \quad (j \in J), \tag{4.2}$$

where I and J are finite index sets, the $\lambda_i^{(j)}$, c_j and d_j are complex algebraic constants, and the t_i are complex variables, together with a polytope $\mathcal{P} \subseteq \mathbb{R}^{2k}$ that is specified by a system of linear inequalities with algebraic coefficients. The problem asks to determine whether there exist $t_1, \ldots, t_k \in \mathbb{C}$ that satisfy the system (4.2) and such that $(\Re(t_1), \ldots, \Re(t_k), \Im(t_1), \ldots, \Im(t_k))$ lies in $\mathcal{P}$.

To establish decidability of the Linear-Exponential Problem, we reduce it to the following *Algebraic-Logarithmic Integer Programming* problem. We remind the reader that a *linear form in logarithms of algebraic numbers* is a number of the form $\beta_0 + \beta_1 \log(\alpha_1) + \cdots + \beta_m \log(\alpha_m)$, where $\beta_0, \alpha_1, \beta_1, \ldots, \alpha_m, \beta_m$ are algebraic numbers and log denotes a fixed branch of the complex logarithm function.

Definition 4.5. An instance of the *Algebraic-Logarithmic Integer Programming Problem* (ALIP) consists of a finite system of equations of the form

$$A\boldsymbol{x} \leq \frac{1}{\pi}\boldsymbol{b},$$

where A is an $m \times n$ matrix with real algebraic entries and where the coordinates of $\boldsymbol{b}$ are real linear forms in logarithms of algebraic numbers. The problem asks to determine whether such a system admits a solution $\boldsymbol{x} \in \mathbb{Z}^n$.

We will also be interested in the following two problems, which will be shown to be undecidable by reduction from MESP.

Definition 4.6. Given square matrices $A_1, \ldots, A_k$ and vectors $\boldsymbol{x}, \boldsymbol{y}$ whose entries are real algebraic numbers and of matching dimensions, the *Generalised Continuous Orbit Problem* (GCOP) consists in deciding whether there exists a matrix C in the multiplicative matrix semigroup generated by the set $\{\exp(A_i t_i), t_i \geq 0, i = 1, \ldots, k\}$ such that $C\boldsymbol{x} = \boldsymbol{y}$. On the other hand, the *Generalised Continuous Skolem Problem* (GCSP) asks whether there exists such a matrix C so that $\boldsymbol{x}^T C \boldsymbol{y} = 0$.

Finally, we revisit the statement of MESP, as MESP is the ultimate object of our analysis. Note that this can be seen as a continuous analogue of the matrix semigroup membership problem, introduced in Section 4.1.1.

Definition 4.1. Given square matrices $A_1, \ldots, A_k$ and C, all of the same dimension, whose entries are real algebraic numbers, the *Matrix-Exponential Semigroup Problem* (MESP) asks whether C a is member of the matrix semigroup generated by

$$\{\exp(A_i t_i) : t_i \in \mathbb{R}, t_i \geq 0, i = 1, \ldots, k\}.$$

4.1.3 Chapter Outline

To help the reader navigate through this chapter, we provide a brief schematic outline of the results/reductions therein. Our undecidability result will be established by the following chain of reductions (where HTP refers to Hilbert's Tenth Problem):

$$\text{HTP} \leq \text{MEP} \leq \text{MESP} \leq \text{GCOP} \leq \text{GCSP}\,.$$

On the other hand, the decidability result in the commutative setting will be established by proving that ALIP is decidable and by the following chain of reductions:

$$\text{Generalised MEP} \leq \text{LEP} \leq \text{ALIP}\,.$$

4.2 Example

In the following example, t_1, t_2 denote fresh variables ranging over the reals. Let $\lambda_1, \lambda_2 \in \mathbb{R} \cap \overline{\mathbb{Q}}$ such that $\lambda_1 > \lambda_2$ and consider the following commuting matrices $A_1, A_2 \in (\mathbb{R} \cap \overline{\mathbb{Q}})^{2\times 2}$:

$$A_i = \begin{pmatrix} \lambda_i & 1 \\ 0 & \lambda_i \end{pmatrix}, i \in \{1, 2\}.$$

One can easily see that

$$\begin{aligned} \exp(A_i t_i) &= \exp(\lambda_i t_i I) \exp(t_i(A_i - \lambda_i I)) \\ &= \exp(\lambda_i t_i) \exp \begin{pmatrix} 0 & t_i \\ 0 & 0 \end{pmatrix} \\ &= \exp(\lambda_i t_i) \begin{pmatrix} 1 & t_i \\ 0 & 1 \end{pmatrix}, i \in \{1, 2\}. \end{aligned}$$

Let $c_1, c_2 \in \mathbb{R} \cap \overline{\mathbb{Q}}$ such that $c_1, c_2 > 0$, and let

$$C = \begin{pmatrix} c_1 & c_2 \\ 0 & c_1 \end{pmatrix}.$$

We would like to determine whether there exists a solution $t_1, t_2 \in \mathbb{R}$, $t_1, t_2 \geq 0$ to

$$\exp(A_1 t_1) \exp(A_2 t_2) = C.$$

This amounts to solving the following system of equations:

$$\begin{cases} \exp(\lambda_1 t_1 + \lambda_2 t_2) = c_1 \\ (t_1 + t_2)\exp(\lambda_1 t_1 + \lambda_2 t_2) = c_2 \end{cases} \Leftrightarrow$$

$$\begin{cases} \exp(t_1(\lambda_1 - \lambda_2) + \frac{c_2}{c_1}\lambda_2) = c_1 \\ t_2 = \frac{c_2}{c_1} - t_1 \end{cases} \Leftrightarrow$$

$$\begin{cases} t_1 = \frac{\log(c_1) - \frac{c_2}{c_1}\lambda_2}{\lambda_1 - \lambda_2} \\ t_2 = \frac{\frac{c_2}{c_1}\lambda_1 - \log(c_1)}{\lambda_1 - \lambda_2} \end{cases}$$

Then $t_1, t_2 \geq 0$ holds if and only if

$$\lambda_2 \leq \frac{c_1}{c_2}\log(c_1) \leq \lambda_1.$$

Whether these inequalities hold amounts to comparing linear forms in logarithms of algebraic numbers. As we will see, this is decidable.

4.3 Decidability in the Commutative Case

We start this section by reducing the Generalised MEP with commuting matrices to LEP . The intuition behind it is quite simple: perform a change of basis so that the matrices $A_1, \ldots, A_k$, as well as C, become block-diagonal matrices, with each block being upper triangular; we can then separate the problem into several sub-instances, corresponding to the diagonal blocks, and finally make use of our uniqueness result concerning strictly upper triangular logarithms of upper unitriangular matrices, established in Theorem 2.3.

Theorem 4.1. *The Generalised MEP with commuting matrices reduces to LEP .*

Proof. Consider an instance of the generalised MEP, as given in Definition 4.3, with commuting $n \times n$ matrices $A_1, \ldots, A_k$ and target matrix C.

We first show how to define a matrix P such that each matrix $P^{-1}A_iP$ is block diagonal, $i = 1, \ldots, k$, with each block being moreover upper triangular.

By Theorem 2.4 we can write $\mathbb{C}^n$ as a direct sum of subspaces $\mathbb{C}^n = \oplus_{j=1}^{b}\mathcal{V}_j$ such that for every subspace $\mathcal{V}_j$ and matrix A_i, $\mathcal{V}_j$ is an invariant subspace of A_i on which A_i has a single eigenvalue $\lambda_i^{(j)}$.

Define a matrix Q by picking an algebraic basis for each $\mathcal{V}_j$ and successively taking the vectors of each basis to be the columns of Q. Then, each matrix $Q^{-1}A_iQ$ is block-diagonal, where the j-th block is a matrix $B_i^{(j)}$ that represents $A_i \restriction \mathcal{V}_j$, $j = 1, \ldots, b$.

Fixing $j \in \{1, \ldots, b\}$, note that the matrices $B_1^{(j)}, \ldots, B_k^{(j)}$ all commute. Thus we may apply Theorem 2.5 to obtain an algebraic matrix M_j such that each matrix $M_j^{-1} B_i^{(j)} M_j$ is upper triangular, $i = 1, \ldots, k$. Thus we can write

$$M_j^{-1} B_i^{(j)} M_j = \lambda_i^{(j)} I + N_i^{(j)}$$

for some strictly upper triangular matrix $N_i^{(j)}$.

We define M to be the block-diagonal matrix with blocks $M_1, \ldots, M_b$. Letting $P = QM$, it is then the case that $P^{-1} A_i P$ is block-diagonal, with the j-th block being $\lambda_i^{(j)} I + N_i^{(j)}$ for $j = 1, \ldots, b$. Now

$$\prod_{i=1}^{k} \exp(A_i t_i) = C \Leftrightarrow \prod_{i=1}^{k} \exp(P^{-1} A_i P t_i) = P^{-1} C P. \tag{4.3}$$

If $P^{-1}CP$ is not block-diagonal, with each block being upper triangular and with the same entries along the diagonal, then Equation (4.3) has no solution and the problem instance must be negative. Otherwise, denoting the blocks $P^{-1}CP$ by $D^{(j)}$ for $j \in \{1, \ldots, b\}$, our problem amounts to simultaneously solving the system of matrix equations

$$\prod_{i=1}^{k} \exp\big((\lambda_i^{(j)} I + N_i^{(j)}) t_i\big) = D^{(j)}, \quad j \in \{1, \ldots, b\} \tag{4.4}$$

with one equation for each block.

For each fixed j, the matrices $N_i^{(j)}$ inherit commutativity from the matrices $B_i^{(j)}$, so we have

$$\begin{aligned} \prod_{i=1}^{k} \exp((\lambda_i^{(j)} I + N_i^{(j)}) t_i) &= \exp\big(\sum_{i=1}^{k} (\lambda_i^{(j)} I + N_i^{(j)}) t_i\big) \\ &= \exp\big(\sum_{i=1}^{k} \lambda_i^{(j)} t_i\big) \cdot \exp\big(\sum_{i=1}^{k} N_i^{(j)} t_i\big). \end{aligned}$$

Hence the system (4.4) is equivalent to

$$\exp\big(\sum_{i=1}^{k} \lambda_i^{(j)} t_i\big) \cdot \exp\big(\sum_{i=1}^{k} N_i^{(j)} t_i\big) = D^{(j)} \tag{4.5}$$

for $j = 1, \ldots, b$.

By assumption, the diagonal entries of each matrix $D^{(j)}$ are equal to a unique value, say $c^{(j)}$. Since the diagonal entries of $\exp\left(\sum_{i=1}^{k} N^{(j)} t_i\right)$ are all 1, the equation

system (4.5) is equivalent to:

$$\exp\big(\sum_{i=1}^{k}\lambda_i^{(j)}t_i\big)=c^{(j)} \text{ and } \exp\big(\sum_{i=1}^{k}N_i^{(j)}t_i\big)=\frac{1}{c^{(j)}}D^{(j)}$$

for $j=1,\dots,b$.

Applying Theorem 2.3, the above system can equivalently be written

$$\exp\big(\sum_{i=1}^{k}\lambda_i^{(j)}t_i\big)=c^{(j)} \text{ and } \sum_{i=1}^{k}N_i^{(j)}t_i=S^{(j)}$$

for some effectively computable matrix $S^{(j)}$ with algebraic entries, $j=1,\dots,b$.

Except for the additional linear equations, this has the form of an instance of LEP. However we can eliminate the linear equations by performing a linear change of variables, i.e., by computing the solution of the system in parametric form. Thus we finally arrive at an instance of LEP. □

In the following result, we essentially solve the system of equations (4.2), reducing it to the simpler problem that really lies at its heart.

Theorem 4.2. *LEP reduces to ALIP.*

Proof. Consider an instance of LEP, comprising a system of equations

$$\exp\left(\sum_{\ell=1}^{k}\lambda_\ell^{(j)}t_\ell\right)=c_j\exp(d_j)\quad j=1,\dots,b, \tag{4.6}$$

and polytope $\mathcal{P}\subseteq\mathbb{R}^{2k}$, as described in Definition 4.4.

Throughout this proof, let log denote a fixed logarithm branch that is defined on all the numbers $c_j,\exp(d_j)$ appearing above, and for which $\log(-1)=i\pi$. Note that if any $c_j=0$ for some j then (4.6) has no solution. Otherwise, by applying log to each equation in (4.6), we get

$$\sum_{\ell=1}^{k}\lambda_\ell^{(j)}t_\ell=d_j+\log(c_j)+2i\pi n_j\quad j=1,\dots,b, \tag{4.7}$$

where $n_j\in\mathbb{Z}$.

The system of equations (4.7) can be written in matrix form as

$$A\boldsymbol{t}\in\boldsymbol{d}+\log(\boldsymbol{c})+2i\pi\mathbb{Z}^b\,,$$

where A is the $b \times k$ matrix with $A_{j,\ell} = \lambda_\ell^{(j)}$ and log is applied pointwise to vectors. Now, defining the convex polytope $\mathcal{Q} \subseteq \mathbb{R}^{2b}$ by

$$\mathcal{Q} = \{(\Re(A\boldsymbol{y}), \Im(A\boldsymbol{y})) : \boldsymbol{y} \in \mathbb{C}^k, (\Re(\boldsymbol{y}), \Im(\boldsymbol{y})) \in \mathcal{P}\},$$

it suffices to decide whether the set $\boldsymbol{d}+\log(\boldsymbol{c})+2i\pi\mathbb{Z}^b$ intersects $\{\boldsymbol{x} \in \mathbb{C}^b : (\Re(\boldsymbol{x}), \Im(\boldsymbol{x})) \in \mathcal{Q}\}$.

Define $f : \mathbb{R}^b \to \mathbb{C}^b$ by $f(\boldsymbol{v}) = \boldsymbol{d} + \log(\boldsymbol{c}) + 2i\pi\boldsymbol{v}$, and define a convex polytope $\mathcal{T} \subseteq \mathbb{R}^b$ by

$$\mathcal{T} = \{\boldsymbol{v} \in \mathbb{R}^b : (\Re(f(\boldsymbol{v})), \Im(f(\boldsymbol{v}))) \in \mathcal{Q}\}.$$

The problem then amounts to deciding whether the convex polytope $\mathcal{T}$ contains an integer point. Crucially, the description of the convex polytope $\mathcal{T}$ is of the form $\pi B\boldsymbol{x} \leq \boldsymbol{b}$, for some matrix B and vector $\boldsymbol{b}$ such that the entries of B are real algebraic and the components of $\boldsymbol{b}$ are real linear forms in logarithms of algebraic numbers. But this is the form of an instance of ALIP . □

We are left with the task of showing that ALIP is decidable. The argument essentially consists of reducing to a lower-dimensional instance whenever possible, and eventually either using the fact that the polytope is bounded to test whether it intersects the integer lattice or using Theorem 2.15 to show that, by a density argument, it must intersect the integer lattice.

Theorem 4.3. *ALIP is decidable.*

Proof. We are given a convex polytope $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d : \pi A\boldsymbol{x} \leq \boldsymbol{b}\}$, where the coordinates $\boldsymbol{b}$ are linear forms in logarithms of algebraic numbers, and need to decide whether this polytope intersects $\mathbb{Z}^d$. Throughout this proof, log denotes the logarithm branch picked at the beginning of the proof of Theorem 4.2. We start by eliminating linear dependencies between the logarithms appearing in the vector $\boldsymbol{b}$, using Masser's Theorem. For example, suppose that

$$b_i = r_0 + r_1 \log(s_1) + \cdots + r_k \log(s_k), \quad r_0, r_1, s_1, \ldots, r_k, s_k \in \overline{\mathbb{Q}}.$$

Due to Theorem 2.11, there exists a non-trivial linear relation with algebraic coefficients amongs $\log(-1), \log(s_1), \ldots, \log(s_k)$ if and only if there is one with integer coefficients. But such integer relations can be computed, since

$$n_0 \log(-1) + n_1 \log(s_1) + \cdots + n_k \log(s_k) = 0 \Leftrightarrow$$
$$(-1)^{n_0} s_1^{n_1} \cdots s_k^{n_k} = 1$$

and since the group of multiplicative relations $\mathcal{L}(-1, s_1, \ldots, s_k)$ can be effectively computed. Whenever $\mathcal{L}(-1, s_1, \ldots, s_k)$ contains a non-zero vector, we use it to eliminate an unnecessary $\log(s_i)$ term, although never eliminating $\log(-1)$. When this process is over, we can see whether each term b_i/π is algebraic or transcendental: it is algebraic if $b_i = \alpha \log(-1), \alpha \in \overline{\mathbb{Q}}$, and transcendental otherwise.

Now, when $\boldsymbol{x} \in \mathbb{Z}^d$, $A\boldsymbol{x}$ is a vector with algebraic coefficients, so whenever b_i/π is transcendental we may alter $\mathcal{P}$ by replacing $\leq$ by $<$ in the i-th inequality, preserving its intersection with $\mathbb{Z}^d$. On the other hand, whenever b_i/π is algebraic, we split our problem into two: in the first one, $\mathcal{P}$ is altered to force equality on the i-th constraint (that is, replacing $\leq$ by $=$), and in the second we force strict inequality (that is, replacing $\leq$ by $<$). We do this for all i, so that no $\leq$ is left in any problem instance, leaving us with finitely many polytopes, each defined by constraints of the form

$$\begin{aligned} K\boldsymbol{x} &= \boldsymbol{k} && (\boldsymbol{k} \in \overline{\mathbb{Q}}^{d_1}) \\ M\boldsymbol{x} &< \boldsymbol{m} && (\boldsymbol{m} \in \overline{\mathbb{Q}}^{d_2}) \\ F\boldsymbol{x} &< \boldsymbol{f} && (\boldsymbol{f} \in \left(\mathbb{R} \setminus \overline{\mathbb{Q}}\right)^{d_3}) \end{aligned}$$

where K, M, F are matrices with algebraic entries. Before proceeding, we eliminate all such empty polytopes; note that emptiness can be decided via Fourier-Motzkin elimination, as shown in Theorem 2.28.

The idea of the next step is to reduce the dimension of all the problem instances at hand until we are left with a number of new instances with full-dimensional open convex polytopes, of the same form as the original one, apart from the fact that all inequalities in their definitions will be strict. To do that, we use the equations $K\boldsymbol{x} = \boldsymbol{k}$ to eliminate variables: note that, whenever there is an integer solution,

$$K\boldsymbol{x} = \boldsymbol{k}, \boldsymbol{x} \in \mathbb{Z}^d \Leftrightarrow \boldsymbol{x} = \boldsymbol{x}_0 + M\boldsymbol{z},$$

where M is a matrix with integer entries, $\boldsymbol{x}_0$ is an integer vector and $\boldsymbol{z}$ ranges over integer vectors over a smaller dimension space, wherein we also define the polytope

$$\mathcal{Q} = \{\boldsymbol{y} : \boldsymbol{x}_0 + M\boldsymbol{y} \in \mathcal{P}\}.$$

Having now eliminated all equality constraints, we are left with a finite set of polytopes of the form $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d : \pi A\boldsymbol{x} < \boldsymbol{b}\}$ that are either empty or full-dimensional and open, and wish to decide whether they intersect the integer lattice of the corresponding space (different instances may lie in spaces of different dimensions, of course). Note that, when $\mathcal{P}$ is non-empty, we can use Fourier-Motzkin elimination

to find a vector $\boldsymbol{q} \in \mathbb{Q}^d$ in its interior, and $\varepsilon > 0$ such that the closed ball of centre $\boldsymbol{q}$ and radius ε with respect to the l_1 norm, which we call $\mathcal{B}$, is contained in $\mathcal{P}$.

The next step is to consider the Minkowski-Weyl decomposition of $\mathcal{P}$, namely $\mathcal{P} = \mathcal{H} + \mathcal{C}$, where $\mathcal{H}$ is the convex hull of finitely many points of $\mathcal{P}$ (which we need not compute) and $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^d : A\boldsymbol{x} \leq \boldsymbol{0}\}$ is a cone with an algebraic description. Note that $\mathcal{P}$ is bounded if and only if $\mathcal{C} = \{\boldsymbol{0}\}$, in which case the problem at hand is simple: consider the polytope $\mathcal{Q}$ with an algebraic description obtained by rounding up each coordinate of $\boldsymbol{b}/\pi$, which has the same conic part as $\mathcal{P}$ and which contains $\mathcal{P}$, and therefore is bounded; finally, compute a bound on $\mathcal{Q}$ (such a bound can be defined in the first-order theory of the reals), which is also a bound on $\mathcal{P}$, and test the integer points within that bound for membership in $\mathcal{P}$. Otherwise,

$$\mathcal{C} = \{\alpha_1 \boldsymbol{c}_1 + \cdots + \alpha_k \boldsymbol{c}_k : \alpha_1, \ldots, \alpha_k \geq 0\},$$

where $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k \in \overline{\mathbb{Q}}^d$ are the extremal rays of $\mathcal{C}$. Note that $\boldsymbol{q} + \mathcal{C} \subseteq \mathcal{P}$ and that $\mathcal{B} + \mathcal{C} \subseteq \mathcal{P}$.

Now we consider a variation of an argument which appears in [51]. Consider the computable set

$$\mathcal{L} = \mathcal{C}^\perp \cap \mathbb{Z}^d .$$

If $\mathcal{L} = \{\boldsymbol{0}\}$ then due to Theorem 2.15 it must be the case that there exist non-negative reals $\lambda_1, \ldots, \lambda_k$ such that

$$\operatorname{dist}\left(\boldsymbol{q} + \sum_{i=1}^{k} \lambda_i \boldsymbol{c}_i, \mathbb{Z}^d\right) \leq \varepsilon,$$

and we know that $\mathcal{P} \cap \mathbb{Z}^d \neq \emptyset$ from the fact that the closed ball $\mathcal{B}$ of centre $\boldsymbol{q}$ and radius ε with respect to the l_1 norm is contained in $\mathcal{P}$.

On the other hand, if $\mathcal{L} \neq \{\boldsymbol{0}\}$, let $\boldsymbol{z} \in \mathcal{L} \setminus \{\boldsymbol{0}\}$. Since $\mathcal{H}$ is a bounded subset of $\mathbb{R}^n$, the set

$$\{\boldsymbol{z}^T \boldsymbol{x} : \boldsymbol{x} \in \mathcal{P}\} = \{\boldsymbol{z}^T \boldsymbol{x} : \boldsymbol{x} \in \mathcal{H}\}$$

is a bounded subset of $\mathbb{R}$. Therefore there exist $a, b \in \mathbb{Z}$ such that

$$\forall \boldsymbol{x} \in \mathcal{P}, a \leq \boldsymbol{z}^T \boldsymbol{x} \leq b,$$

so we can reduce our problem to $b - a + 1$ smaller-dimensional instances by finding the integer points of $\{\boldsymbol{x} \in \mathcal{P} : \boldsymbol{z}^T \boldsymbol{x} = i\}$, for $i \in \{a, \ldots, b\}$. Note that we have seen earlier in the proof how to reduce the dimension of the ambient space when the polytope $\mathcal{P}$ is contained in an affine hyperplane. $\square$

4.4 Undecidability of the Non-Commutative Case

In this section we show that the Matrix-Exponential Problem is undecidable in the case of non-commuting matrices. We show undecidability for the most fundamental variant of the problem, as given in Definition 4.3, in which the matrices have real entries and the variables t_i range over the non-negative reals. Recall that this problem is decidable in the commutative case by the results of the previous section.

4.4.1 Matrix-Exponential Problem with Constraints

The proof of undecidability in the non-commutative case is by reduction from Hilbert's Tenth Problem. The reduction proceeds via several intermediate problems. These problems are obtained by augmenting MEP with various classes of arithmetic constraints on the real variables that appear in the statement of the problem.

Definition 4.7. We consider the following three classes of arithmetic constraints over real variables $t_1, t_2, \ldots$:

- $\mathcal{E}_{\pi\mathbb{Z}}$ comprises constraints of the form $t_i \in \alpha + \beta\pi\mathbb{Z}$, where α and $\beta \neq 0$ are real-valued constants such that $\cos(2\alpha\beta^{-1})$, β are both algebraic numbers.
- $\mathcal{E}_{+}$ comprises linear equations of the form $\alpha_1 t_1 + \cdots + \alpha_n t_n = \alpha_0$, for $\alpha_0, \ldots, \alpha_n$ real algebraic constants.
- $\mathcal{E}_{\times}$ comprises equations of the form $t_\ell = t_i t_j$.

A class of constraints $\mathcal{E} \subseteq \mathcal{E}_{\pi\mathbb{Z}} \cup \mathcal{E}_{+} \cup \mathcal{E}_{\times}$ induces a generalisation of the MEP problem as follows:

Definition 4.8 (MEP with Constraints)**.** Given a class of constraints $\mathcal{E} \subseteq \mathcal{E}_{\pi\mathbb{Z}} \cup \mathcal{E}_{+} \cup \mathcal{E}_{\times}$, the problem MEP($\mathcal{E}$) is as follows. An instance consists of real algebraic matrices $A_1, \ldots, A_k, C$ and a finite set of constraints $E \subseteq \mathcal{E}$ on real variables $t_1, \ldots, t_k$. The question is whether there exist non-negative real values for $t_1, \ldots, t_k$ such that $\prod_{i=1}^{k} e^{A_i t_i} = C$ and the constraints E are all satisfied.

Note that in the above definition of MEP($\mathcal{E}$) the set of constraints E only mentions real variables $t_1, \ldots, t_k$ appearing in the matrix equation $\prod_{i=1}^{k} e^{A_i t_i} = C$. However, without loss of generality, we can allow constraints to mention fresh variables t_i, for $i > k$, since we can always define a corresponding matrix $A_i = 0$ for such variables for then $e^{A_i t_i} = I$ has no effect on the matrix product. In other words, we can without loss of generality have constraints in $\mathcal{E}$ with existentially quantified variables. In particular, we have the following useful observations:

- We can express inequality constraints of the form $t_i \neq \alpha$ in $\mathcal{E}_+ \cup \mathcal{E}_\times$ by using fresh variables t_j, t_ℓ. Indeed $t_i \neq \alpha$ is satisfied whenever there exist values of t_j and t_ℓ such that $t_i = t_j + \alpha$ and $t_j t_\ell = 1$.
- By using fresh variables, $\mathcal{E}_+ \cup \mathcal{E}_\times$ can express polynomial constraints of the form $P(t_1, \ldots, t_n) = t$ for P a polynomial with integer coefficients.

We illustrate the above two observations in an example.

Example. Consider the problem, given matrices A_1, A_2 and C, to decide whether there exist $t_1, t_2 \geq 0$ such that

$$e^{A_1 t_1} e^{A_2 t_2} = C \text{ and } t_1^2 - 1 = t_2, t_2 \neq 0\,.$$

This is equivalent to the following instance of MEP($\mathcal{E}_+ \cup \mathcal{E}_\times$): decide whether there exist $t_1, \ldots, t_5 \geq 0$ such that

$$\prod_{i=1}^{5} e^{A_i t_i} = C \text{ and } t_1 t_1 = t_3, t_3 - 1 = t_2, t_2 t_4 = t_5, t_5 = 1$$

where A_1, A_2 and C are as above and $A_3 = A_4 = A_5 = 0$.

We will make heavy use of the following proposition to combine several instances of the constrained MEP into a single instance by combining matrices block-wise.

Proposition 4.4. *Given real algebraic matrices $A_1, \ldots, A_k, C$ and $A'_1, \ldots, A'_k, C'$, there exist real algebraic matrices $A''_1, \ldots, A''_k$, C'' such that for all $t_1, \ldots, t_k$:*

$$\prod_{i=1}^{k} e^{A''_i t_i} = C'' \qquad \Leftrightarrow \qquad \left(\prod_{i=1}^{k} e^{A_i t_i} = C\right) \wedge \left(\prod_{i=1}^{k} e^{A'_i t_i} = C'\right).$$

Proof. For any $i \in \{1, \ldots, k\}$, define

$$A''_i = \begin{bmatrix} A_i & 0 \\ 0 & A'_i \end{bmatrix}, \qquad C'' = \begin{bmatrix} C & 0 \\ 0 & C' \end{bmatrix}.$$

The result follows because the matrix exponential can be computed block-wise (as is clear from its power series definition):

$$\prod_{i=1}^{k} e^{A''_i t_i} = \prod_{i=1}^{k} \begin{bmatrix} e^{A_i t_i} & 0 \\ 0 & e^{A'_i t_i} \end{bmatrix} = \begin{bmatrix} \prod_{i=1}^{k} e^{A_i t_i} & 0 \\ 0 & \prod_{i=1}^{k} e^{A'_i t_i} \end{bmatrix}.$$

□

We remark that in the statement of Proposition 4.4 the two matrix equations that are combined are over the same set of variables. However, we can clearly combine any two matrix equations for which the common variables appear in the same order in the respective products.

The core of the reduction is to show that the constraints in $\mathcal{E}_{\pi\mathbb{Z}}, \mathcal{E}_+$ and $\mathcal{E}_\times$ do not make the MEP problem harder: one can always encode them using the matrix product equation.

Proposition 4.5. *MEP*$(\mathcal{E}_{\pi\mathbb{Z}} \cup \mathcal{E}_+ \cup \mathcal{E}_\times)$ *reduces to MEP*$(\mathcal{E}_+ \cup \mathcal{E}_\times)$.

Proof. Let $A_1, \ldots, A_k, C$ be real algebraic matrices and $E \subseteq \mathcal{E}_{\pi\mathbb{Z}} \cup \mathcal{E}_+ \cup \mathcal{E}_\times$ a finite set of constraints on real variables $t_1, \ldots, t_k$. Since E is finite it suffices to show how to eliminate a single constraint in $\mathcal{E}_{\pi\mathbb{Z}} \cap E$ from E.

Let $t_j \in \alpha + \beta\pi\mathbb{Z}$ be a constraint in E. By definition of $\mathcal{E}_{\pi\mathbb{Z}}$ we have that $\cos(2\alpha\beta^{-1}), \sin(2\alpha\beta^{-1})$ and $\beta \neq 0$ are real algebraic. Now define the following extra matrices:

$$A'_j = \begin{bmatrix} 0 & 2\beta^{-1} \\ -2\beta^{-1} & 0 \end{bmatrix}, C' = \begin{bmatrix} \cos(2\alpha\beta^{-1}) & \sin(2\alpha\beta^{-1}) \\ -\sin(2\alpha\beta^{-1}) & \cos(2\alpha\beta^{-1}) \end{bmatrix}.$$

Our assumptions ensure that A'_j and C' are both real algebraic.

We now have the following chain of equivalences:

$$\begin{aligned} e^{A'_j t_j} = C' &\Leftrightarrow \begin{bmatrix} \cos(2t_j\beta^{-1}) & \sin(2t_j\beta^{-1}) \\ -\sin(2t_j\beta^{-1}) & \cos(2t_j\beta^{-1}) \end{bmatrix} = C' \\ &\Leftrightarrow \cos(2t_j\beta^{-1}) = \cos(2\alpha\beta^{-1}) \\ &\qquad \wedge \sin(2t_j\beta^{-1}) = \sin(2\alpha\beta^{-1}) \\ &\Leftrightarrow 2\beta^{-1}t_j \equiv 2\alpha\beta^{-1} \mod 2\pi \\ &\Leftrightarrow t_j \in \alpha + \beta\pi\mathbb{Z}. \end{aligned}$$

Thus the additional matrix equation $e^{A'_j t_j} = C'$ is equivalent to the constraint $t_j \in \alpha + \beta\pi\mathbb{Z}$. Applying Proposition 4.4 we can thus eliminate this constraint. □

Proposition 4.6. *MEP*$(\mathcal{E}_+ \cup \mathcal{E}_\times)$ *reduces to MEP*$(\mathcal{E}_+)$.

Proof. Let $A_1, \ldots, A_k, C$ be real algebraic matrices and $E \subseteq \mathcal{E}_+ \cup \mathcal{E}_\times$ a finite set of constraints on variables $t_1, \ldots, t_k$. We proceed as above, showing how to remove each constraint in $\mathcal{E}_\times$ from E. In so doing we potentially increase the number of matrices and add new constraints from $\mathcal{E}_+$.

Let $t_l = t_i t_j$ be an equation in E. To eliminate this equation the first step is to introduce fresh variables x, x', y, y', z and add the constraints

$$t_i = x,\, t_j = y,\, t_\ell = z,$$

which are all in $\mathcal{E}_+$. We now add a new matrix equation over the fresh variables x, x', y, y', z that is equivalent to the constraint $xy = z$. Since this matrix equation involves a new set of variables we are free to the set the order of the matrix products, which is crucial to express the desired constraint.

The key gadget is the following matrix product equation, which holds for any $x, x', y, y', z \geqslant 0$:

$$\begin{bmatrix} 1 & 0 & -z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -y' \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & x & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -x' & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & x - x' & z - xy \\ 0 & 1 & y - y' \\ 0 & 0 & 1 \end{bmatrix}.$$

Notice that each of the matrices on the left-hand side of the above equation has a single non-zero off-diagonal entry. Crucially each matrix of this form can be expressed as an exponential. Indeed we can write the above equation as a matrix-exponential product

$$e^{B_1 z} e^{B_2 y'} e^{B_3 x} e^{B_4 y} e^{B_5 x'} = \begin{bmatrix} 1 & x - x' & z - xy \\ 0 & 1 & y - y' \\ 0 & 0 & 1 \end{bmatrix}$$

for matrices

$$B_1 = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad B_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad B_5 = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$B_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \qquad B_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Thus the constraint $xy = z$ can be expressed as

$$e^{B_1 z} e^{B_2 y'} e^{B_3 x} e^{B_4 y} e^{B_5 x'} = I\,. \tag{4.8}$$

Again, we can apply Proposition 4.4 to combine the equation (4.8) with the matrix equation from the original problem instance and thus encode the constraint $z = xy$.

□

Proposition 4.7. *MEP($\mathcal{E}_+$) reduces to MEP.*

Proof. Let $A_1, \ldots, A_k, C$ be real algebraic matrices and $E \subseteq \mathcal{E}_+$ a set of constraints. We proceed as above, showing how to eliminate each constraint from E that lies in $\mathcal{E}_+$, while preserving the set of solutions of the instance.

Let $\beta + \sum_{i=1}^{k} \alpha_i t_i = 0$ be an equation in E. Recall that $\beta, \alpha_1, \ldots, \alpha_k$ are real algebraic. Define the extra matrices $A'_1, \ldots, A'_k$ and C' as follows:

$$A'_i = \begin{bmatrix} 0 & \alpha_i \\ 0 & 0 \end{bmatrix}, \qquad C' = \begin{bmatrix} 1 & -\beta \\ 0 & 1 \end{bmatrix}.$$

Our assumptions ensure that $A'_1, \ldots, A'_k$ and C' are all real algebraic. Furthermore, the following extra product equation becomes:

$$\begin{aligned} \prod_{i=1}^{k} e^{A'_i t_i} = C \Leftrightarrow \prod_{i=1}^{k} \begin{bmatrix} 1 & \alpha_i t_i \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} 1 & -\beta \\ 0 & 1 \end{bmatrix} \\ \Leftrightarrow \sum_{i=1}^{k} \alpha_i t_i &= -\beta\,. \end{aligned}$$

□

Combining Proposition 4.5, Proposition 4.6, and Proposition 4.7 we have:

Proposition 4.8. *MEP*$(\mathcal{E}_{\pi\mathbb{Z}} \cup \mathcal{E}_+ \cup \mathcal{E}_\times)$ *reduces to MEP.*

4.4.2 Reduction from Hilbert's Tenth Problem

Theorem 4.9. *MEP is undecidable in the non-commutative case.*

Proof. We have seen in the previous section that the problem MEP$(\mathcal{E}_{\pi\mathbb{Z}} \cup \mathcal{E}_+ \cup \mathcal{E}_\times)$ reduces to MEP without constraints. Thus it suffices to reduce Hilbert's Tenth Problem to MEP$(\mathcal{E}_{\pi\mathbb{Z}} \cup \mathcal{E}_+ \cup \mathcal{E}_\times)$. In fact the matrix equation will not play a role in the target of this reduction, only the additional constraints!

Let P be a polynomial of total degree d in k variables with integer coefficients. From P we build a homogeneous polynomial Q, by adding a new variable λ:

$$Q(\mathbf{x}, \lambda) = \lambda^d P\left(\frac{x_1}{\lambda}, \ldots, \frac{x_k}{\lambda}\right).$$

Note that Q still has integer coefficients. Furthermore, we have the relationship

$$Q(\mathbf{x}, 1) = P(\mathbf{x}).$$

As we have seen previously, it is easy to encode Q with constraints, in the sense that we can compute a finite set of constraints $E_Q \subseteq \mathcal{E}_+ \cup \mathcal{E}_\times$ mentioning variables

$t_0, \ldots, t_m, \lambda$ such that E is satisfied if and only if $t_0 = Q(t_1, \ldots, t_k, \lambda)$. Note that E_Q may need to mention variables other than $t_1, \ldots, t_k$ to do that. Another finite set of equations $E_{\pi\mathbb{Z}} \subseteq \mathcal{E}_{\pi\mathbb{Z}}$ is used to encode that $t_1, \ldots, t_k, \lambda \in \pi\mathbb{Z}$. Finally, $E_= \subseteq \mathcal{E}_+ \cup \mathcal{E}_\times$ is used to encode $t_0 = 0$ and $1 \leqslant \lambda \leqslant 4$. The latter is done by adding the polynomial equations $\lambda = 1 + \alpha^2$ and $\lambda = 4 - \beta^2$ for some α and β. Finally we have the following chain of equivalences:

$$\begin{aligned}
&\exists t_0, \ldots, \lambda \geqslant 0 \text{ s.t. } E_Q \cup E_{\pi\mathbb{Z}} \cup E_= \text{ is satisfied} \\
&\quad \Leftrightarrow \exists t_1, \ldots, \lambda \geqslant 0 \text{ s.t. } 0 = Q(t_1, \ldots, t_k, \lambda) \\
&\qquad \wedge t_1, \ldots, t_k, \lambda \in \pi\mathbb{Z} \wedge 1 \leqslant \lambda \leqslant 4 \\
&\quad \Leftrightarrow \exists n_1, \ldots, n_k \in \mathbb{N} \text{ s.t. } 0 = Q(\pi n_1, \ldots, \pi n_k, \pi) \\
&\quad \Leftrightarrow \exists n_1, \ldots, n_k \in \mathbb{N} \text{ s.t. } 0 = \pi^d Q(n_1, \ldots, n_k, 1) \\
&\quad \Leftrightarrow \exists n_1, \ldots, n_k \in \mathbb{N} \text{ s.t. } 0 = P(n_1, \ldots, n_k).
\end{aligned}$$

□

4.4.3 Enforcing a matrix product order

In this section, we will present a gadget matrix-exponential semigroup that can enforce a certain partial order on the matrices reaching a particular target. This will be useful to establish the reduction MEP $\leq$ MESP. More precisely, we will exhibit five matrices W, X, Y, Z and G such that any product $G = \prod_{i=1}^{p} e^{A_i t_i}$ where $t_i > 0$ and $A_i \in \{W, X, Y, Z\}$ is such that all the "X" appear before the "Y". Define the following matrices:

$$W = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \qquad X = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad Z = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix},$$

$$Y = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

One easily computes the exponentials W, X, Y, Z:

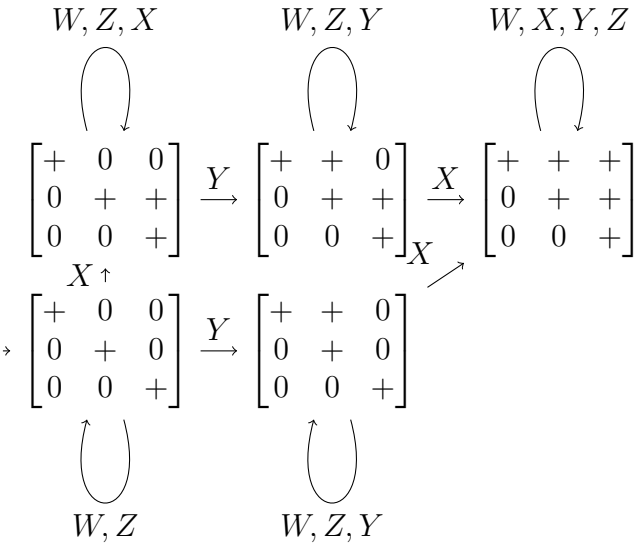
$$e^{Wt} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^t & 0 \\ 0 & 0 & e^{2t} \end{bmatrix}, e^{Xt} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}, e^{Yt} = \begin{bmatrix} 1 & t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, e^{Zt} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-t} & 0 \\ 0 & 0 & e^{-2t} \end{bmatrix}.$$

The crux of the proof will be based on the following asymmetry between X and Y, which leaves the top right corner zero in one case but nonzero in the other. As we will later observe, once the top right corner is nonzero, it cannot be cleared.

$$e^{Xt}e^{Yu} = \begin{bmatrix} 1 & u & 0 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}, \qquad e^{Yt}e^{Xu} = \begin{bmatrix} 1 & u & tu \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}.$$

Proposition 4.10. *If there exist $p \in \mathbb{N}$, $A_i \in \{W, X, Y, Z\}$ and $t_i > 0$, for $i \in \{1, \dots, p\}$, such that $\prod_{i=1}^{p} e^{A_i t_i} = G$, then the product contains at least one[2] "X" and one "Y", and all the "X" appear before the "Y". Formally, there exist i and j such that $A_i = X$ and $A_j = Y$, and for any such choice we have $i < j$.*

Proof. First observe that any such product is always an upper triangular matrix with non-negative entries (because all the matrices have non-negative entries) and positive entries on the diagonal. Let M be such a matrix; we denote its coefficients by 0 if they are zero and + if they are positive. The following automaton should be read as follows: an arrow from M to M' annotated with A means that any product of a matrix of the shape of M by e^{At} with $t > 0$ will give a matrix with the shape of M'. Note that the empty product gives the identity. One easily checks that the following transitions hold.



Starting from the identity and applying the different products $e^{A_i t_i}$ in the automaton, it is clear that the only way to reach a matrix of the shape of G is to have all the "X" before "Y". Formally, by contradiction, if there were $i < j$ such that $A_i = Y$

[2]Note that this is not entirely trivial because we required only positive t_i in the product.

and $A_j = X$ then by the automaton, we would end up with a matrix where the top right corner is nonzero, which contradicts $G = \prod_{i=1}^{p} e^{A_i t_i}$. □

The previous lemma shows that this semigroup enforces a partial order on the matrices in products that reach the matrix G. The next lemma shows that G can indeed be reached using this kind of products, essentially proving that G belongs to this semigroup.

Proposition 4.11. *For any positive real t, there exists a non-negative real u such that*

$$e^{Wu}e^{Xt}e^{Yt}e^{Zu} = G \quad \text{or} \quad e^{Zu}e^{Xt}e^{Yt}e^{Wu} = G.$$

Proof. Consider the following products for an arbitrary $u \geqslant 0$:

$$\begin{aligned} e^{Wu}e^{Xt}e^{Yt}e^{Zu} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{u} & 0 \\ 0 & 0 & e^{2u} \end{bmatrix} \begin{bmatrix} 1 & t & 0 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-u} & 0 \\ 0 & 0 & e^{-2u} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{u} & 0 \\ 0 & 0 & e^{2u} \end{bmatrix} \begin{bmatrix} 1 & te^{-u} & 0 \\ 0 & e^{-u} & te^{-2u} \\ 0 & 0 & e^{-2u} \end{bmatrix} \\ &= \begin{bmatrix} 1 & te^{-u} & 0 \\ 0 & 1 & te^{-u} \\ 0 & 0 & 1 \end{bmatrix}, \end{aligned}$$

and

$$e^{Zu}e^{Xt}e^{Yt}e^{Wu} = \begin{bmatrix} 1 & te^{u} & 0 \\ 0 & 1 & te^{u} \\ 0 & 0 & 1 \end{bmatrix}.$$

If $t \geqslant 1$ then choosing $u = \ln t \geqslant 0$ in the first product gives G, otherwise choosing $u = \ln \frac{1}{t} \geqslant 0$ in the second product gives G. □

4.4.4 Undecidability of the semigroup problem

We will now show the undecidability of MESP. First, we remind the reader of Definition 4.1.

Definition 4.1. Given square matrices $A_1, \ldots, A_k$ and C, all of the same dimension, whose entries are real algebraic numbers, the *Matrix-Exponential Semigroup Problem* (MESP) asks whether C a is member of the matrix semigroup generated by

$$\{\exp(A_i t_i) : t_i \in \mathbb{R}, t_i \geq 0, i = 1, \ldots, k\}.$$

In short, the difference is that we do not fix the order of the matrices in the product, and that each matrix may be used more than once. We will now show the following key result.

Theorem 4.12. *MESP is undecidable in the non-commutative case.*

Proof. We have seen in the previous section that MEP is undecidable in the non-commutative case. Thus it suffices to reduce MEP to MESP to show undecidability.

Let $A_1, \ldots, A_k, C \in \overline{\mathbb{Q}}^{n\times n}$ be an instance of MEP. Denote by I_m the identity of size m, 0_m the zero matrix of size m. Recall the 3×3 matrices W, X, Y, Z, G, defined in Section 4.4.3. For every $i \in \{2, \ldots, k-1\}$, define the following matrices:

$$B_i = \begin{bmatrix} A_i & & & & \\ & 0_{3(i-2)} & & & \\ & & Y & & \\ & & & X & \\ & & & & 0_{3(k-1-i)} \end{bmatrix}, \quad B'_i = \begin{bmatrix} 0_n & & & & \\ & 0_{3(i-2)} & & & \\ & & Y & & \\ & & & X & \\ & & & & 0_{3(k-1-i)} \end{bmatrix}.$$

We also define the matrices matrices

$$B_1 = \begin{bmatrix} A_1 & & & \\ & X & & \\ & & 0_{3(k-2)} & \end{bmatrix}, \qquad B'_1 = \begin{bmatrix} 0_n & & & \\ & X & & \\ & & 0_{3(k-2)} & \end{bmatrix},$$

$$B_k = \begin{bmatrix} A_1 & & & \\ & 0_{3(k-2)} & & \\ & & Y & \end{bmatrix}, \qquad B'_k = \begin{bmatrix} 0_n & & & \\ & 0_{3(k-2)} & & \\ & & Y & \end{bmatrix},$$

and, for every $i \in \{1, \ldots, k-1\}$,

$$W_i = \begin{bmatrix} 0_n & & & & \\ & 0_{3(i-1)} & & & \\ & & W & & \\ & & & 0_{3(k-1-i)} \end{bmatrix}, \qquad Z_i = \begin{bmatrix} 0_n & & & & \\ & 0_{3(i-1)} & & & \\ & & Z & & \\ & & & 0_{3(k-1-i)} \end{bmatrix}.$$

Finally, we define the target matrix:

$$C' = \begin{bmatrix} C & & & \\ & G & & \\ & & \ddots & \\ & & & G \end{bmatrix}.$$

We can now define our MESP instance as follows, the target matrix is C' and the semigroup $\mathcal{G}$ is generated by

$$\left\{ e^{B_i t_i}, e^{B'_i t_i} : t_i \geqslant 0, i = 1, \ldots, k \right\}$$
$$\cup \left\{ e^{W_i t_i}, e^{Z_i t_i} : t_i \geqslant 0, i = 1, \ldots, k-1 \right\}.$$

We claim the original instance of MEP is satisfiable if and only if $C' \in \mathcal{G}$. Let us examine both direction independently.

Assume that the MEP instance is satisfiable. Then there exist $t_1, \ldots, t_k \geqslant 0$ such that:

$$\prod_{i=1}^{k} e^{A_i t_i} = C.$$

Define $\tau = \max\{t_1, \ldots, t_k\} + 1$ (note that $\tau > 0$) and $t'_i = \tau - t_i \geqslant 0$ for every $i \in \{1, \ldots, k\}$. A straightforward calculation shows that:

$$\prod_{i=1}^{k} \left(e^{B_i t_i} e^{B'_i t'_i} \right) = \begin{bmatrix} \prod_{i=1}^{k} e^{A_i t_i} & & & \\ & e^{X\tau} e^{Y\tau} & & \\ & & \ddots & \\ & & & e^{X\tau} e^{Y\tau} \end{bmatrix} = \begin{bmatrix} C & & & \\ & U & & \\ & & \ddots & \\ & & & U \end{bmatrix},$$

where $U = e^{X\tau} e^{Y\tau}$. Apply Proposition 4.11 to get $\lambda \geqslant 0$ such that either $e^{W\lambda} U e^{Z\lambda} = G$ or $e^{Z\lambda} U e^{W\lambda} = G$. In the first case, conclude by checking that

$$\prod_{i=1}^{k-1} e^{W_i \lambda} \prod_{i=1}^{k} \left(e^{B_i t_i} e^{B'_i t'_i} \right) \prod_{i=1}^{k-1} e^{Z_i \lambda} = \begin{bmatrix} C & & & \\ & G & & \\ & & \ddots & \\ & & & G \end{bmatrix} = C'.$$

In the second case, exchange the W_i and Z_i to get same result. This concludes the proof that the MESP instance is satisfiable, since all the products belong to $\mathcal{G}$.

Assume that the MESP instance is satisfiable. Then there exists $t_1, \ldots, t_m > 0$ (we can always take them positive) and $M_1, \ldots, M_m \in \left\{ B_i, B'_i : i = 1, \ldots, k \right\} \cup \left\{ W_i, Z_i : i = 1, \ldots, k-1 \right\}$ such that

$$\prod_{j=1}^{m} e^{M_j t_j} = C'. \tag{4.9}$$

Observe that by construction, this product has the following form:

$$\prod_{j=1}^{m} e^{M_j t_j} = \begin{bmatrix} V & & & \\ & U_1 & & \\ & & \ddots & \\ & & & U_{k-1} \end{bmatrix},$$

where V belongs to the semigroup generated by $\{e^{A_i t} : t \geqslant 0\}$ and U_i belongs to the semigroup generated by $\{e^{Wt}, e^{Xt}, e^{Yt}, e^{Zt} : t \geqslant 0\}$. Since (4.9) implies that $U_i = G$, we can apply Proposition 4.10 to get each product producing U_i must have all its "X" before its "Y". Furthermore, each U_i must contain at least one X and one Y in its product. For any $i \in \{1, \ldots, k\}$, let k_i (resp., k_i') denote the first (resp., last) index j such that $M_j = B_i$ or B_i'. Those indices exists because of the proposition since at least one B_i or B_i' must appear for every i to get both an X and a Y in each product giving U_i. Obviously $k_i \leqslant k_i'$ by definition. We now claim that the proposition implies that:

$$k_1 \leqslant k_1' < k_2 \leqslant k_2' < k_3 \cdots < k_{k-1} \leqslant k_{k-1}'.$$

Indeed, Proposition 4.10 ensures that in the product giving U_1, all the "X" appear before the "Y", but the only matrices that contribute some X to U_1 are B_1 and B_1', and the only matrices that contribute some Y to U_1 are B_2 and B_2'. Thus $k_1' < k_2$, i.e. the last "X" coming from B_1 or B_1' is before the first "Y" coming from B_2 or B_2'. A similar reasoning ensures that $k_2' < k_3$ and so on. This shows that for any $i \in \{1, \ldots, k\}$, if $M_j = B_i$ then $j \in \{k_i, \ldots, k_i'\}$. Thus all the B_1 appear before the B_2 which appear before the B_3 and so on. However, since the B_i are the only ones to contribute to V, then V must be of the form:

$$V = \prod_{i=1}^{k} e^{A_i t_i'},$$

where $t_i' \geqslant 0$ is the sum of all t_j such that $M_j = B_i$. Finally $V = C$, so the instance of MEP is satisfiable.

□

4.5 Generalised Continuous Orbit and Skolem Problems

We start by recalling the definition of the Generalised Continuous Orbit and Skolem Problems.

Definition 4.6. Given square matrices $A_1, \ldots, A_k$ and vectors $\boldsymbol{x}, \boldsymbol{y}$ whose entries are real algebraic numbers and of matching dimensions, the *Generalised Continuous Orbit Problem* (GCOP) consists in deciding whether there exists a matrix C in the multiplicative matrix semigroup generated by the set $\{\exp(A_i t_i), t_i \geq 0, i = 1, \ldots, k\}$ such that $C\boldsymbol{x} = \boldsymbol{y}$. On the other hand, the *Generalised Continuous Skolem Problem* (GCSP) asks whether there exists such a matrix C so that $\boldsymbol{x}^T C \boldsymbol{y} = 0$.

We will now prove that these are both undecidable.

Theorem 4.13. *The Generalised Continuous Orbit Problem is undecidable.*

Proof. This can be shown by reduction from MESP. In particular, given square matrices $B_1, \ldots, B_k, C$, we construct matrices $A_1, \ldots, A_k$ and vectors $\boldsymbol{x}, \boldsymbol{y}$ for which

$$\prod_{i=1}^{k} \exp(B_i t_i) = C \Leftrightarrow \prod_{i=1}^{k} \exp(A_i t_i)\boldsymbol{x} = \boldsymbol{y}. \tag{4.10}$$

Let $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ be the columns of C, from left to right, and let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ denote the canonical basis of $\mathbb{R}^n$. Then (4.10) can be achieved by setting, for each $i \in \{1, \ldots, n\}$,

$$A_i = \begin{pmatrix} B_i & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_i \end{pmatrix},$$

as well as

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_n \end{pmatrix} \text{ and } \boldsymbol{y} = \begin{pmatrix} \boldsymbol{c}_1 \\ \vdots \\ \boldsymbol{c}_n \end{pmatrix}.$$

□

Finally, we show the last result of the present chapter.

Theorem 4.14. *The Generalised Continuous Skolem Problem is undecidable.*

Proof. This can be shown by reduction from the Generalised Continuous Orbit Problem. Similarly to what we did in the previous proof, we will define matrices $C_1, \ldots, C_k$ and vectors $\boldsymbol{w}, \boldsymbol{z}$ such that

$$\prod_{i=1}^{k} \exp(A_i t_i)\boldsymbol{x} = \boldsymbol{y} \Leftrightarrow \boldsymbol{w}^T \prod_{i=1}^{k} \exp(C_i t_i)\boldsymbol{z} = 0\,.$$

Let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ denote the canonical basis of $\mathbb{R}^n$. Moreover, let

$$B_i = \begin{pmatrix} A_i & 0 \\ 0 & 0 \end{pmatrix}, \boldsymbol{u}_j = \begin{pmatrix} \boldsymbol{e}_j \\ -\boldsymbol{e}_j \end{pmatrix}, \boldsymbol{v} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$$

and

$$C_i = \begin{pmatrix} B_i \otimes I + I \otimes B_i & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_i \otimes I + I \otimes B_i \end{pmatrix}.$$

Then

$$\begin{aligned}
&\prod_{i=1}^{k} \exp(A_i t_i)\boldsymbol{x} = \boldsymbol{y} \\
\Leftrightarrow &\sum_{j=1}^{n} \left(\boldsymbol{u}_j^T \prod_{i=1}^{k} \exp(B_i t_i)\boldsymbol{v} \right)^2 = 0 \\
\Leftrightarrow &\sum_{j=1}^{n} \left((\boldsymbol{u}_j \otimes \boldsymbol{u}_j)^T \prod_{i=1}^{k} \left(\exp(B_i t_i) \otimes \exp(B_i t_i)\right) (\boldsymbol{v} \otimes \boldsymbol{v}) \right) = 0 \\
\Leftrightarrow &\sum_{j=1}^{n} \left((\boldsymbol{u}_j \otimes \boldsymbol{u}_j)^T \prod_{i=1}^{k} \exp\left((B_i \otimes I + I \otimes B_i) t_i\right) (\boldsymbol{v} \otimes \boldsymbol{v}) \right) = 0 \\
\Leftrightarrow &\begin{pmatrix} (\boldsymbol{u}_1 \otimes \boldsymbol{u}_1)^T \\ \vdots \\ (\boldsymbol{u}_n \otimes \boldsymbol{u}_n)^T \end{pmatrix}^T \prod_{i=1}^{k} \exp(C_i t_i) \begin{pmatrix} \boldsymbol{v} \otimes \boldsymbol{v} \\ \vdots \\ \boldsymbol{v} \otimes \boldsymbol{v} \end{pmatrix} = 0.
\end{aligned}$$

The result then follows by taking

$$\boldsymbol{w} = \begin{pmatrix} (\boldsymbol{u}_1 \otimes \boldsymbol{u}_1)^T \\ \vdots \\ (\boldsymbol{u}_n \otimes \boldsymbol{u}_n)^T \end{pmatrix} \text{ and } \boldsymbol{z} = \begin{pmatrix} \boldsymbol{v} \otimes \boldsymbol{v} \\ \vdots \\ \boldsymbol{v} \otimes \boldsymbol{v} \end{pmatrix}.$$

□

4.6 Turing-degree of MESP

We are interested in classifying the Turing-degree of MESP. We will place MESP in the first level of the arithmetical hierarchy conditionally on Schanuel's conjecture being true (Theorem 4.16), and in the second level unconditionally (Theorem 4.15).

Given a set A, we define $A^{(0)}$ to be A, and for each $n \geq 1$ we inductively define $A^{(n+1)}$ as the Halting Problem for Turing machines with access to an oracle for solving $A^{(n)}$. We call $A^{(n)}$ the n-th *Turing jump* of A.

We have already seen that $\emptyset^{(1)} \leq MESP$ holds.

Theorem 4.15. *$MESP \leq \emptyset^{(2)}$.*

Proof. Consider the functions

$$f_{\boldsymbol{w}}(\boldsymbol{t}) = \left\| \prod_{i=1}^{|\boldsymbol{w}|} \exp(A_{w_i} t_i) - C \right\|_2^2, \ \boldsymbol{w} \in \{1, \ldots, k\}^\star$$

and note that each $f_{\boldsymbol{w}}$ is an exponential-polynomial which has the non-negative reals as co-domain. Clearly, this instance of MESP is positive if and only if some $f_{\boldsymbol{w}}$ has a tangential zero.

Before proceeding, note that exponential-polynomials are closed under differentiation, and that they are computable functions.

Let $\boldsymbol{b} : \mathbb{N} \to \{1, \ldots, k\}^\star$ be any computable surjection.

For each $n \in \mathbb{N}$ and $\boldsymbol{w} \in \{\boldsymbol{b}(1), \ldots, \boldsymbol{b}(n)\}$, consider the Turing machine $A_{\boldsymbol{w},n}$ which does the following: for each $m \in \mathbb{N}$, partition $[0, n]^{|\boldsymbol{w}|}$ in an uniform grid, with mesh size m^{-1}, and compute the approximate value of $f_{\boldsymbol{w}}$ with error at most m^{-1} at each grid point; if it is ever the case that all the approximate values of $f_{\boldsymbol{w}}$ are greater than $\left(1 + \frac{L_{\boldsymbol{w},n}\sqrt{|\boldsymbol{w}|}}{2}\right) m^{-1}$ (where $L_{\boldsymbol{w},n}$ is an upper bound on $\|\nabla f_{\boldsymbol{w}}\| \restriction_{[0,n]}$, which we can compute by using the triangle inequality and the monotonicity of the exponential function), $A_{\boldsymbol{w},n}$ halts. Due to the Mean Value Theorem and to the compactness of $[0, n]$, $A_{\boldsymbol{w},n}$ halts if and only if $f_{\boldsymbol{w}} \restriction_{[0,n]}$ does not have zeroes. Thus, the instance of MESP is positive if and only if some $A_{\boldsymbol{w},n}$ loops.

Now, consider the Turing machine B with access to a Halting Problem oracle (that is, a $\emptyset^{(1)}$ oracle) which, for each $n \in \mathbb{N}$ and $\boldsymbol{w} \in \{\boldsymbol{b}(1), \ldots, \boldsymbol{b}(n)\}$, uses the oracle to decide whether $A_{\boldsymbol{w},n}$ halts; if that is ever not the case, B halts. Finally, note that B halts if and only if the MESP instance in consideration is positive. □

Moreover, the following result holds.

Theorem 4.16. *If Schanuel's conjecture is true, MESP* $\leq \emptyset^{(1)}$.

Proof. Let $f_{\boldsymbol{w}}, \boldsymbol{w} \in \{1, \ldots, k\}^\star$ be as in the proof of Theorem 4.15. Consider the Turing Machine T which, for each $n \in \mathbb{N}$ and $\boldsymbol{w} \in \{\boldsymbol{b}(1), \ldots, \boldsymbol{b}(n)\}$, uses Theorem 2.21 to decide whether $f_{\boldsymbol{w}}$ admits a zero in the region $[0, n]^{|\boldsymbol{w}|}$, and halts when such a zero is found. Then T halts if and only if the instance of MESP under consideration is positive. □

4.7 Conclusion

We have shown that the Matrix-Exponential Problem is undecidable in general, but decidable when the matrices $A_1, \ldots, A_k$ commute. We have also showed that the Matrix-Exponential Semigroup Problem is undecidable, by designing a gadget to enforce an order in the products, and derived the undecidability of the generalised versions of the Continuous Orbit and Continuous Skolem Problems to a multi-matrix

setting. This is analogous to what was known for the discrete version of this problem, in which the matrix exponentials e^{At} are replaced by matrix powers A^n. Finally, we show that these problems are in Σ_1 if Schanuel's conjecture is true, and in Σ_2 unconditionally. It would be interesting to show membership in Σ_1 unconditionally. In the discrete version of the problem, membership in Σ_1 follows trivially from the countability of the space of candidate solutions.

It would be interesting to look at possibly decidable restrictions of the MEP/MESP, for example the case where $k = 2$ with a non-commuting pair of matrices, which was shown to be decidable for the discrete analogue of this problem in [11]. Bounding the dimension of the ambient vector space could also yield decidability, which has been partly accomplished in the discrete case in [75].

Finally, deriving upper and lower bounds for the computational complexity of the commutative case of this problem would also be a worthwhile task.

Chapter 5

The Polytope Escape Problem

5.1 Introduction

In ambient space $\mathbb{R}^d$, a *continuous linear dynamical system* is a trajectory $\boldsymbol{x}(t)$, where t ranges over the non-negative reals, defined by a differential equation $\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t))$ in which the function f is *affine* or *linear*. If the initial point $\boldsymbol{x}(0)$ is given, the differential equation uniquely defines the entire trajectory. (Linear) dynamical systems have been extensively studied in mathematics, physics, and engineering, and more recently have played an increasingly important role in computer science, notably in the modelling and analysis of cyber-physical systems; a recent and authoritative textbook on the matter is [3].

In the study of dynamical systems, particularly from the perspective of control theory, considerable attention has been given to the study of *invariant sets*, i.e., subsets of $\mathbb{R}^d$ from which no trajectory can escape; see, e.g., [26, 16, 5, 81]. Our focus in the present chapter is on sets with the dual property that *no trajectory remains trapped*. Such sets play a key role in analysing *liveness* properties in cyber-physical systems (see, for instance, [3]): discrete progress is ensured by guaranteeing that all trajectories (i.e., from any initial starting point) must eventually reach a point at which they 'escape' (temporarily or permanently) the set in question.

More precisely, given an affine function $f : \mathbb{R}^d \to \mathbb{R}^d$ and a convex polytope $\mathcal{P} \subseteq \mathbb{R}^d$, both specified using rational coefficients encoded in binary, we consider the *Polytope Escape Problem*, which asks whether there is some point $\boldsymbol{x}_0$ in $\mathcal{P}$ for which the corresponding trajectory of the solution to the differential equation

$$\begin{cases} \dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases}$$

is entirely contained in $\mathcal{P}$. Our main result is to show that this problem is decidable by reducing it in polynomial time to the decision version of linear programming with real algebraic coefficients, which itself reduces in polynomial time to deciding the truth of a sentence in the first-order theory of the reals, a problem whose complexity is known to lie between **NP** and **PSPACE** [24]. Our algorithm makes use of spectral techniques and relies, among others, on tools from Diophantine approximation.

It is interesting to note that a seemingly closely related problem, that of determining whether a given trajectory of a linear dynamical system ever hits a given hyperplane (also known as the *Continuous Skolem Problem*), is not known to be decidable; see, in particular, [12, 34, 33]. When the target is instead taken to be a single point (rather than a hyperplane), the corresponding reachability question (known as the *Continuous Orbit Problem*) can be decided in polynomial time [41].

5.2 Main Results

The Polytope Escape Problem for continuous linear dynamical systems consists of deciding, given an affine function $f : \mathbb{R}^d \to \mathbb{R}^d$ and a convex polytope $\mathcal{P} \subseteq \mathbb{R}^d$, whether there exists an initial point $\boldsymbol{x}_0 \in \mathcal{P}$ for which the trajectory of the unique solution to the differential equation $\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t)), \boldsymbol{x}(0) = \boldsymbol{x}_0, t \geq 0$, is entirely contained in $\mathcal{P}$. A starting point $\boldsymbol{x}_0 \in \mathcal{P}$ is said to be *trapped* if the trajectory of the corresponding solution is contained in $\mathcal{P}$, and *eventually trapped* if the trajectory of the corresponding solution contains a trapped point. Therefore, the Polytope Escape Problem amounts to deciding whether a trapped point exists, which in turn is equivalent to deciding whether an eventually trapped point exists.

The goal of this section is to prove the following result:

Theorem 5.1. *The Polytope Escape Problem is polynomial-time reducible to the decision version of linear programming with algebraic coefficients.*

A d-dimensional instance of the Polytope Escape Problem is a pair $(f, \mathcal{P})$, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is an affine function and $\mathcal{P} \subseteq \mathbb{R}^d$ is a convex polytope. In this formulation we assume that all numbers involved in the definition of f and $\mathcal{P}$ are rational.[1]

[1]The assumption of rationality is required to justify some of our complexity claims (e.g., Jordan canonical forms are only known to be polynomial-time computable for matrices with rational coordinates). Nevertheless, our procedure remains valid in a more general setting, and in fact, the overall $\exists\mathbb{R}$ complexity of our algorithm would not be affected if one allowed real algebraic numbers when defining problem instances.

An instance $(f, \mathcal{P})$ of the Polytope Escape Problem is said to be *homogeneous* if f is a linear function and $\mathcal{P}$ is a convex polytope cone (in particular, $\boldsymbol{x} \in \mathcal{P}, \alpha > 0 \Rightarrow \alpha\boldsymbol{x} \in \mathcal{P}$).

The restriction of the Polytope Escape Problem to homogeneous instances is called the homogeneous Polytope Escape Problem.

Lemma 5.2. *The Polytope Escape Problem is polynomial-time reducible to the homogeneous Polytope Escape Problem.*

Proof. Let $(f, \mathcal{P})$ be an instance of the Polytope Escape Problem in $\mathbb{R}^d$, and write

$$f(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{a} \text{ and } \mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d : B_1\boldsymbol{x} > \boldsymbol{b}_1 \wedge B_2\boldsymbol{x} \geq \boldsymbol{b}_2\}.$$

Now define

$$A' = \begin{pmatrix} A & \boldsymbol{a} \\ \mathbf{0}^T & 0 \end{pmatrix}, B_1' = \begin{pmatrix} B_1 & -\boldsymbol{b}_1 \\ \mathbf{0}^T & 1 \end{pmatrix}, B_2' = \begin{pmatrix} B_2 & -\boldsymbol{b}_2 \end{pmatrix},$$
$$\mathcal{P}' = \left\{ \begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix} \in \mathbb{R}^{d+1} : B_1' \begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix} > \mathbf{0} \wedge B_2' \begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix} \geq \mathbf{0} \right\},$$

and

$$g \begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix} = A' \begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix}.$$

Then $(g, \mathcal{P}')$ is a homogeneous instance of the Polytope Escape Problem.

It is clear that $\boldsymbol{x}(t)$ satisfies the differential equation $\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t))$ if and only if $\begin{pmatrix} \boldsymbol{x}(t) \\ 1 \end{pmatrix}$ satisfies the differential equation $\begin{pmatrix} \dot{\boldsymbol{x}} \\ \dot{y} \end{pmatrix} = g \begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix}$. In general, in any trajectory $\begin{pmatrix} \boldsymbol{x} \\ y \end{pmatrix}$ that satisfies this last differential equation, the y-component must be constant.

We claim that $(f, \mathcal{P})$ is a positive instance of the Polytope Escape Problem if and only if $(g, \mathcal{P}')$ is a positive instance. Indeed, if the point $\boldsymbol{x}_0 \in \mathbb{R}^d$ is trapped in $(f, \mathcal{P})$ then the point $\begin{pmatrix} \boldsymbol{x}_0 \\ 1 \end{pmatrix}$ is trapped in $(g, \mathcal{P}')$. Conversely, suppose that $\begin{pmatrix} \boldsymbol{x}_0 \\ y_0 \end{pmatrix}$ is trapped in $(g, \mathcal{P}')$. Then, since $B_1' \begin{pmatrix} \boldsymbol{x}_0 \\ y_0 \end{pmatrix} > \mathbf{0}$, we must have $y_0 > 0$. Scaling, it follows that $\begin{pmatrix} y_0^{-1}\boldsymbol{x}_0 \\ 1 \end{pmatrix}$ is also trapped in $(g, \mathcal{P}')$. This implies that $y_0^{-1}\boldsymbol{x}_0$ is trapped in $(f, \mathcal{P})$. □

We remind the reader that the unique solution to the differential equation $\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t)), \boldsymbol{x}(0) = \boldsymbol{x}_0, t \geq 0$, where $f(\boldsymbol{x}) = A\boldsymbol{x}$, is given by $\boldsymbol{x}(t) = \exp(At)\boldsymbol{x}_0$.

In this setting, the sets of trapped and eventually trapped points are, respectively,

$$T = \{\boldsymbol{x}_0 \in \mathbb{R}^d : \forall t \geq 0, \exp(At)\boldsymbol{x}_0 \in \mathcal{P}\}$$
$$ET = \{\boldsymbol{x}_0 \in \mathbb{R}^d : \exists t \geq 0, \exp(At)\boldsymbol{x}_0 \in T\}.$$

Note that both T and ET are convex subsets of $\mathbb{R}^d$.

Lemma 5.3. *The homogeneous Polytope Escape Problem is polynomial-time reducible to the decision version of linear programming with algebraic coefficients.*

Proof. Let $\boldsymbol{x}_0 = \boldsymbol{x}_0^r + \boldsymbol{x}_0^c$, where $\boldsymbol{x}_0^r \in \mathcal{V}^r$ and $\boldsymbol{x}_0^c \in \mathcal{V}^c$. We start by showing that if $\boldsymbol{x}_0$ lies in the set T of trapped points then its component $\boldsymbol{x}_0^r$ in the real eigenspace $\mathcal{V}^r$ lies in the set ET of eventually trapped points. Due to the fact that the intersection of finitely many convex polytopes is still a convex polytope, it suffices to prove this claim for the case when $\mathcal{P}$ is defined by a single inequality—say $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{b}^T\boldsymbol{x} \triangleright 0\}$, where $\triangleright$ is either $>$ or $\geq$.

We may assume that $\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^c$ is not identically zero, as in that case

$$\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0 \equiv \boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^r$$

and our claim holds trivially. Also, if $\boldsymbol{x}_0 \in T$, it cannot hold that $\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^r \equiv 0$, since $\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^c$ is negative infinitely often by Corollary 2.32.

Suppose that $\boldsymbol{x}_0 \in T$ and let (ρ, m) and (η, j) be the dominant indices for $\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^r$ and $\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^c$ respectively. Then by Proposition 2.2 we have

$$\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^r = \exp(\rho t)t^m(c + o(1)) \tag{5.1}$$

as $t \to \infty$, where c is a non-zero real number. We will show that $c > 0$, from which it follows that $\boldsymbol{x}_0^r \in ET$.

It must hold that $(\eta, j) \preceq (\rho, m)$. Indeed, if $(\eta, j) \succ (\rho, m)$, then, as $t \to \infty$,

$$\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0 = \exp(\eta t)t^j \left(\underbrace{\frac{\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^c}{\exp(\eta t)t^j}}_{\text{A}} + o(1) \right),$$

but the limit inferior of the term A above is strictly negative by Corollary 2.32, contradicting the fact that $\boldsymbol{x}_0 \in T$.

If $(\eta, j) = (\rho, m)$, then, as $t \to \infty$,

$$\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0 = \exp(\rho t)t^m \left(c + \frac{\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^c}{\exp(\rho t)t^m} + o(1) \right),$$

and by invoking Corollary 2.32 as above, it follows that $c > 0$.

Finally, if $(\eta, j) \prec (\rho, m)$, then, as $t \to \infty$,

$$\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0^c = \exp(\rho t)t^m \cdot o(1)\,, \tag{5.2}$$

and hence, by (5.1) and (5.2), it follows that

$$\boldsymbol{b}^T \exp(At)\boldsymbol{x}_0 = \exp(\rho t)t^m \left(c + o(1)\right) .$$

From the fact that $\boldsymbol{x}_0 \in T$ and that $c \neq 0$ we must have $c > 0$.

In all cases it holds that $c > 0$ and hence $\boldsymbol{x}_0^r \in ET$.

Having argued that $ET \neq \emptyset$ iff $ET \cap \mathcal{V}^r \neq \emptyset$, we will now show that the set $ET \cap \mathcal{V}^r$ is a convex polytope that we can efficiently compute. As before, it suffices to prove this claim for the case when $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{b}^T\boldsymbol{x} \triangleright 0\}$ (where $\triangleright$ is either $>$ or $\geq$).

In what follows, we let $[K]$ denote the set $\{0, \ldots, K-1\}$. We can write

$$\boldsymbol{b}^T \exp(At) = \sum_{(\eta,j)\in\sigma(A)\times[\nu(A)]} \exp(\eta t)t^j \boldsymbol{u}_{(\eta,j)}^T ,$$

where $\boldsymbol{u}_{(\eta,j)}^T$ is the vector of coefficients of $t^j \exp(\eta t)$ in $\boldsymbol{b}^T \exp(At)$.

Note that if $\boldsymbol{x} \in \mathcal{V}^r$ and $(\eta, j) \in (\sigma(A) \setminus \mathbb{R}) \times \mathbb{N}_0$, then $\boldsymbol{u}_{(\eta,j)}^T\boldsymbol{x} = 0$, as $\boldsymbol{u}_{(\eta,j)}^T\boldsymbol{x}$ is the coefficient of $t^j \exp(\eta t)$ in $\boldsymbol{b}^T \exp(At)\boldsymbol{x}$, and $\mathcal{V}^r$ is invariant under $\exp(At)$. Moreover,

$$ET \cap \mathcal{V}^r = (\mathcal{B} \cap \mathcal{C}) \cup \begin{cases} \{\boldsymbol{0}\} & \text{if } \triangleright \text{ is } \geq; \\ \emptyset & \text{if } \triangleright \text{ is } >; \end{cases}$$

where

$$\begin{aligned} \mathcal{B} &= \bigcap_{(\eta,j)\in(\sigma(A)\setminus\mathbb{R})\times[\nu(A)]} \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{u}_{(\eta,j)}^T\boldsymbol{x} = 0\} \\ \mathcal{C} &= \bigcup_{(\eta,j)\in(\sigma(A)\cap\mathbb{R})\times[\nu(A)]} \Bigg[\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{u}_{(\eta,j)}^T\boldsymbol{x} > 0\} \cap \\ &\qquad \bigcap_{(\rho,m)\succ(\eta,j)} \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{u}_{(\rho,m)}^T\boldsymbol{x} = 0\}\Bigg]. \end{aligned}$$

The set $ET \cap \mathcal{V}^r$ can be seen to be convex from the above characterisation. Alternatively, note that ET can be shown to be convex from its definition and that $\mathcal{V}^r$ is convex, therefore so must be their intersection. Thus $ET \cap \mathcal{V}^r$ must be a convex polytope whose definition possibly involves canonically-represented real algebraic numbers, and the Polytope Escape Problem reduces to testing this polytope for non-emptiness. □

5.3 Conclusion

We have shown that the Polytope Escape Problem for continuous-time linear dynamical systems is decidable, and in fact, polynomial-time reducible to the decision problem for the existential theory of real closed fields. Given an instance of the problem $(f, \mathcal{P})$, with f an affine map, our decision procedure involves analysing the real eigenstructure of the linear operator $g(\boldsymbol{x}) := f(\boldsymbol{x}) - f(\mathbf{0})$. In fact, we showed that all complex eigenvalues could essentially be ignored for the purposes of deciding this problem.

Interestingly, the seemingly closely related question of whether a given single trajectory of a linear dynamical system remains trapped within a given polytope, on the other hand, appears to be considerably more challenging and is not known to be decidable. In that instance, it seems that the influence of the complex eigenstructure cannot simply be discarded.

Chapter 6

Reachability for Linear Time-Invariant Control Systems

6.1 Introduction

In this chapter, we study a basic problem in control theory, namely the point-to-point controllability problem for both continuous- and discrete-time linear time-invariant (henceforth LTI) systems.

A discrete-time LTI system $(\boldsymbol{x}_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^d$ with control set $\mathcal{U} \subseteq \mathbb{R}^d$ satisfies the evolution rule $\boldsymbol{x}_{n+1} = A\boldsymbol{x}_n + \boldsymbol{u}_n$, where A is a $d \times d$ matrix and $\boldsymbol{u}_n \in \mathcal{U}$. In words, the next state $\boldsymbol{x}_{n+1}$ of the system is obtained by applying a time-invariant linear function to the previous state $\boldsymbol{x}_n$ and adding a control $\boldsymbol{u}_n$ from a time-invariant set $\mathcal{U}$ of controls.

Given two points $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d$, the *point-to-point controllability question for discrete-time LTI systems* consists in deciding whether there exist $n \in \mathbb{N}$ and $(\boldsymbol{u}_i)_{i=0}^{n-1} \subseteq \mathcal{U}$ such that $\boldsymbol{x}_0 = \boldsymbol{s}$ and $\boldsymbol{x}_n = \boldsymbol{t}$.

We show that this problem is undecidable when $\mathcal{U}$ is non-convex (in particular, when it is a finite union of convex polytopes) and that there is a reduction from Skolem's Problem when $\mathcal{U}$ is a convex polytope. This contrasts with the case where $\mathcal{U}$ is a linear subspace, in which case the problem reduces to the Orbit Problem, known to be decidable [46].

Similarly to the discrete case, a continuous-time LTI system $\boldsymbol{x}(t) \subseteq \mathbb{R}^d$ with control set $\mathcal{U} \subseteq \mathbb{R}^d$ is a function satisfying the evolution rule $\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + \boldsymbol{u}(t)$ where A is a $d\times d$ matrix and $\boldsymbol{u}$ is a measurable function with codomain $\mathcal{U}$. The *point-to-point controllability question for continuous-time LTI systems* consists in deciding whether there exist $t \geq 0$ and $\boldsymbol{u} : [0, t] \to \mathcal{U}$ such that $\boldsymbol{x}(0) = \boldsymbol{s}$ and $\boldsymbol{x}(t) = \boldsymbol{t}$.

We show that this problem is in **PTIME** when $\mathcal{U}$ is a linear subspace, by presenting a polynomial-time reduction to the Continuous Orbit Problem, which is known to be in **PTIME** [28].

A survey of computational complexity results in control theory can be found in [16].

6.2 Discrete-time systems

6.2.1 Hard Instances

In this section, we present a hardness result for the point-to-point controllability problem for LTI systems where the set of admissible controls is a compact convex polytope.

We recall that Skolem's problem and the Positivity Problem are two long-standing open problems, whose decidability has not yet been determined. It is known that the Positivity Problem is Skolem-hard [69]. Instead of reducing from the Positivity Problem directly, we shall proceed by reducing from the following problem, which has been shown to be Positivity-hard in [2]:

Definition 6.1. Given a column-stochastic matrix $M \in \mathbb{Q}^{d\times d}$ and a number $r \in \mathbb{Q} \cap [0,1]$, the *Markov Reachability Problem* consists in determining whether there exists a number $n \in \mathbb{N}$ such that $(M^n)_{1,2} \geq r$.

We will now show the following result:

Theorem 6.1. *The point-to-point controllability problem for LTI systems whose set of admissible controls are compact convex polytopes is Positivity-hard.*

Proof. Given a column-stochastic matrix $M \in \mathbb{Q}^{d\times d}$ and a number $r \in \mathbb{Q} \cap [0,1]$, we define the matrix $A = \mathrm{diag}(M, 0, 0, 1) \in \mathbb{Q}^{(d+3)\times(d+3)}$ and the compact convex polytope

$$\mathcal{P} = \{(-\boldsymbol{x}, y, \boldsymbol{x}\cdot\mathbf{1}, \boldsymbol{x}\cdot\mathbf{1}), \boldsymbol{x} \geq \mathbf{0}, \boldsymbol{x}\cdot\mathbf{1} \leq 1, 0 \leq y \leq \boldsymbol{x}\cdot\boldsymbol{e}_1\} \subseteq \mathbb{R}^{d+3},$$

as well as the source $\boldsymbol{s} = (\boldsymbol{e}_2, 0, 0, 0) \in \mathbb{Q}^{d+3}$ and target $\boldsymbol{t} = (\mathbf{0}, r, 1, 1) \in \mathbb{Q}^{d+3}$.

First, suppose that there exists $n \in \mathbb{N}$ such that $(M^n)_{1,2} \geq r$. Consider the sequence of controls $(\boldsymbol{u}_i)_{i=0}^{i=n-1} \subseteq \mathcal{P}$ given by $\boldsymbol{u}_0 = \cdots = \boldsymbol{u_{n-2}} = \mathbf{0}$ and $\boldsymbol{u}_{n-1}$ is the element from $\mathcal{P}$ corresponding to $\boldsymbol{x} = M^n\boldsymbol{e}_2$ and $y = r$. This sequence clearly controls $\boldsymbol{s}$ to $\boldsymbol{t}$.

On the other hand, suppose there exists a sequence $(\boldsymbol{u}_i)_{i=0}^{i=n-1}$ controlling $\boldsymbol{s}$ to $\boldsymbol{t}$. From the fact that the coordinates $d+2$ and $d+3$ of $\boldsymbol{t}$ are equal, and noting that multiplying by the matrix A erases coordinate $d+2$ but not $d+3$, it follows that $\boldsymbol{u}_{n-1}$ is the only non-zero control, that is, $\boldsymbol{u}_0 = \cdots = \boldsymbol{u}_{n-2} = \boldsymbol{0}$. Therefore, at time $n-1$, we will be in state $(M^{n-1}\boldsymbol{e}_2, 0, 0, 0)$, and the only way to reach $\boldsymbol{t}$ in the remaining step is to take $\boldsymbol{u}_{n-1} \in \mathcal{P}$ with $\boldsymbol{x} = M^n\boldsymbol{e}_2$ and $y = r$, otherwise one of the first d coordinates will be non-zero. It then follows that $(M^n)_{1,2} \geq r$, by looking at coordinate $d+1$. □

6.2.2 Encoding Hilbert's Tenth Problem

Given $k+1$ invertible matrices $A_1, \ldots, A_k, C \in \mathbb{Q}^{d\times d}$, the *Generalised Matrix Powering Problem for invertible matrices* consists in deciding whether there exist $n_1, \ldots, n_k \in \mathbb{Z}$ such that

$$\prod_{i=1}^{k} A_i^{n_i} = C.$$

The techniques used to show the next result are very similar to those employed in Section 4.4. This result is a strengthening of [11], which we will need in the proof of Theorem 6.4.

Theorem 6.2. *The Generalised Matrix Powering Problem for invertible matrices is undecidable.*

Proof. We will show this result by reducing from Hilbert's Tenth Problem. Given a polynomial $p \in \mathbb{Z}[n_1, \ldots, n_k]$, it is easy to express $p(n_1, \ldots, n_k)$ as a conjunction of relations of the following form (noting that we may need to introduce fresh variables):

- $z = k$, where $k \in \mathbb{Z}$
- $z = x + y$
- $z = xy$.

We start by showing how to encode each of these constraints as an instance of the Generalised Matrix Powering Problem for invertible matrices. Firstly, note that

$$z = k \Leftrightarrow \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^z = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}.$$

Secondly, note that

$$z = x + y \Leftrightarrow \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^x \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^y \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}^z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Thirdly, note that

$$z = xy \Leftrightarrow \exists x', y' \in \mathbb{Z}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & x - x' & z - xy \\ 0 & 1 & y - y' \\ 0 & 0 & 1 \end{pmatrix}$$

and that the latter matrix is just equal to

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{z} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}^{y'} \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{x} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}^{y} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{x'}.$$

Finally, conjunction can be achieved by making use of separate matrix blocks:

$$\prod_{i=1}^{k} A_i^{n_i} = C \wedge \prod_{i=1}^{k} B_i^{n_i} = D \Leftrightarrow \prod_{i=1}^{k} \begin{pmatrix} A_i & 0 \\ 0 & B_i \end{pmatrix}^{n_i} = \begin{pmatrix} C & 0 \\ 0 & D \end{pmatrix}.$$

□

Definition 6.2. Given invertible matrices $A_1, \ldots, A_k \in \mathbb{Q}^{d \times d}$ and two non-zero vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Q}^d$, the *vector reachability problem for invertible matrices* consists in deciding whether there exist $n_1, \ldots, n_k \in \mathbb{Z}$ such that

$$\prod_{i=1}^{k} A_i^{n_i} \boldsymbol{x} = \boldsymbol{y}.$$

Note that this problem is similar to the Generalised Orbit Problem, which was shown to be decidable when $A_1, \ldots, A_k \in SL(2, \mathbb{Z})$ [74], but here we fix the order in which we multiply the matrices $A_1, \ldots, A_k$.

Theorem 6.3. *The vector reachability problem for invertible matrices is undecidable.*

Proof. This can be shown by reduction from the generalised matrix powering problem for invertible matrices. In particular, given invertible matrices $A_1, \ldots, A_k, B \in \mathbb{Q}^{d \times d}$, letting $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d$ denote the columns of B, and letting $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ denote the canonical basis of R^d, the result follows from the fact that

$$\prod_{i=1}^{k} A_i^{n_i} = B \Leftrightarrow \prod_{i=1}^{k} \begin{pmatrix} A_i & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A_i \end{pmatrix}^{n_i} \begin{pmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_d \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_d \end{pmatrix}.$$

□

6.2.3 Undecidable instances

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

Theorem 6.4. *The point-to-point controllability problem for LTI systems whose sets of admissible controls are disjoint unions of finitely many closed convex polytopes is undecidable.*

Proof. We prove this by reduction from the vector reachability problem for invertible matrices (defined in Section 6.2.2).

Let $A_1, \dots, A_k \in \mathbb{Q}^{d\times d}$ be invertible matrices and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{Q}^d$ be non-zero vectors.

For each $i \in \{1, \dots, k\}$, we define

$$B_i = \begin{pmatrix} I_d & 0 & 0 \\ 0 & A_i & 0 \\ 0 & 0 & A_i^{-1} \end{pmatrix}$$

and $M = \mathrm{diag}(B_k, \dots, B_1, I_d, 1) \in \mathbb{Q}^{((3k+1)d+1)\times((3k+1)d+1)}$.

Moreover, we let, for each $i \in \{1, \dots, k\}$,

$$\begin{aligned}
\mathcal{P}_i^{(1)} &= \{\mathbf{0}\}^{3(k-i)d} \times \{(-\boldsymbol{z}, \boldsymbol{z}, \mathbf{0}, \mathbf{0}), \boldsymbol{z} \in \mathbb{R}^d\} \times \{\mathbf{0}\}^{3(i-1)d} \times \{1\} \subseteq \mathbb{R}^{(3k+1)d+1} \\
\mathcal{P}_i^{(2)} &= \{\mathbf{0}\}^{3(k-i)d} \times \{(-\boldsymbol{z}, \mathbf{0}, \boldsymbol{z}, \mathbf{0}), \boldsymbol{z} \in \mathbb{R}^d\} \times \{\mathbf{0}\}^{3(i-1)d} \times \{1\} \subseteq \mathbb{R}^{(3k+1)d+1} \\
\mathcal{Q}_i^{(1)} &= \{\mathbf{0}\}^{3(k-i)d} \times \{(\mathbf{0}, -\boldsymbol{z}, \mathbf{0}, \boldsymbol{z}), \boldsymbol{z} \in \mathbb{R}^d\} \times \{\mathbf{0}\}^{3(i-1)d} \times \{1\} \subseteq \mathbb{R}^{(3k+1)d+1} \\
\mathcal{Q}_i^{(2)} &= \{\mathbf{0}\}^{3(k-i)d} \times \{(\mathbf{0}, \mathbf{0}, -\boldsymbol{z}, \boldsymbol{z}), \boldsymbol{z} \in \mathbb{R}^d\} \times \{\mathbf{0}\}^{3(i-1)d} \times \{1\} \subseteq \mathbb{R}^{(3k+1)d+1}.
\end{aligned}$$

We also let $\mathcal{P}_i = \mathcal{P}_i^{(1)} \dot\cup \mathcal{P}_i^{(2)}$ and $\mathcal{Q}_i = \mathcal{Q}_i^{(1)} \dot\cup \mathcal{Q}_i^{(2)}$. We define the set of admissible controls as

$$\mathcal{U} = \bigcup_{i=1}^{k} (\mathcal{P}_i \dot\cup \mathcal{Q}_i) \dot\cup \{\mathbf{0}\}.$$

Finally, defining $\boldsymbol{s} = (\boldsymbol{x}, \mathbf{0}, \dots, \mathbf{0}, 0)$, and $\boldsymbol{t} = (\mathbf{0}, \dots, \mathbf{0}, \boldsymbol{y}, 2k)$, it holds that $\boldsymbol{s}$ can be controlled to $\boldsymbol{t}$ if and only if there exist $n_1, \dots, n_k \in \mathbb{Z} \setminus \{0\}$ such that

$$\prod_{i=1}^{k} A_i^{n_i} \boldsymbol{x} = \boldsymbol{y}.$$

For the "if" implication, suppose that such $n_1, \dots, n_k$ exist. We can use precisely $2k$ non-zero controls, which will correspond to times $t_1 = 0, t_2 = |n_1|, t_3 = |n_1| + 1, \dots, t_{2k-1} = |n_1| + \cdots + |n_{k-1}| + k - 1, t_{2k} = |n_1| + \cdots + |n_k| + k - 1$. For each

$i \in \{0, \ldots, k-1\}$, $\boldsymbol{u}_{t_{2i+1}}$ will belong $\mathcal{P}_{k-i}^{(1)}$ or $\mathcal{P}_{k-i}^{(2)}$ when $n_{k-i} \geq 0$ and $n_{k-i} < 0$ respectively, with

$$\boldsymbol{z} = \prod_{j=k-i+1}^{k} A_j^{n_j} \boldsymbol{x}.$$

On the other hand, for each $i \in \{0, \ldots, k-1\}$, $\boldsymbol{u}_{t_{2i+2}}$ will belong to $\mathcal{Q}_{k-i}^{(1)}$ or $\mathcal{Q}_{k-i}^{(2)}$ when $n_{k-i} \geq 0$ and $n_{k-i} < 0$ respectively, with

$$\boldsymbol{z} = \prod_{j=k-i}^{k} A_j^{n_j} \boldsymbol{x}.$$

It is clear that this sequence controls $\boldsymbol{s}$ to $\boldsymbol{t}$.

We now proceed to showing the more challenging "only if" implication. We first note that, for any sequence controlling $\boldsymbol{s}$ to $\boldsymbol{t}$, for all $i \in \{1, \ldots, k\}$, there exists $r \in \{1, 2\}$ such that both $\mathcal{P}_i^{(r)}$ and $\mathcal{Q}_i^{(r)}$ are used exactly once each and that both $\mathcal{P}_i^{(3-r)}$ and $\mathcal{Q}_i^{(3-r)}$ are never used. This claim is easy to prove: $\boldsymbol{s}$ has a non-zero component in the first block (of dimension d), whilst $\boldsymbol{t}$ does not, and so we need to use an element of $\mathcal{P}_1$ to make that component $\mathbf{0}$. As a result of that, either the second or third block will have a non-zero component, which needs to be cleared in order to hit $\boldsymbol{t}$, and therefore an element of $\mathcal{Q}_1$ needs to be used. Afterwards, we will have a non-zero component in the fourth block, which again needs to be cleared (unless $k = 1$, of course). The same argument can then be applied inductively. That each $\mathcal{P}_i$ and each $\mathcal{Q}_i$ can only be used once follows from the fact that we can only use $2k$ controls over all, and that each needs to be used at least once.

We suppose, for notational simplicity and without loss of generality (as A_i and A_i^{-1} can be exchanged), that all used controls come from the $\mathcal{P}_i^{(1)}$ and $\mathcal{Q}_i^{(1)}$. We also suppose that these controls are taken with $\boldsymbol{z} = \boldsymbol{u}_i$ and $\boldsymbol{z} = \boldsymbol{v}_i$ respectively (the reader may refer to the definition of the control set for clarifying this notation). After the controls are applied (whatever their order), for some $n_1, m_1, \ldots, n_k, m_k \in \mathbb{N}$ [1], we will be the following state:

$$\left(-\boldsymbol{u}_k, A_k^{n_k}\boldsymbol{u}_k - A_k^{m_k}\boldsymbol{v}_k, \mathbf{0}, \boldsymbol{v}_k - \boldsymbol{u}_{k-1}, \ldots, \boldsymbol{v}_2 - \boldsymbol{u}_1, A_1^{n_1}\boldsymbol{u}_1 - A_1^{m_1}\boldsymbol{v}_1, \mathbf{0}, \boldsymbol{v}_1, 2k\right).$$

For this to be equal to $\boldsymbol{t}$, the following needs to hold:

$$\boldsymbol{y} = \boldsymbol{v}_1 = A_1^{n_1-m_1}\boldsymbol{u}_1 = A_1^{n_1-m_1}\boldsymbol{v}_2 = A_1^{n_1-m_1}A_2^{n_2-m_2}\boldsymbol{u}_2 = \cdots = \prod_{i=1}^{k} A_i^{n_i-m_i}\boldsymbol{x}.$$

This concludes our proof. □

[1] Note that n_i corresponds to the number of steps since a control from $\mathcal{P}_i$ was applied, whilst m_i corresponds to the number of steps since a control from $\mathcal{Q}_i$ was applied.

6.3 Continuous-time systems

6.3.1 Decidable Instances: Reducing to the Continuous Orbit Problem

Given two matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times d}$, consider the following differential equation:

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t). \tag{6.1}$$

We recall that the continuous point-to-point controllability problem for linear time-invariant (LTI) systems amounts to deciding whether, given initial and target points $s, t \in \mathbb{R}^n$, there exists a real $T \geq 0$ and a measurable function $\boldsymbol{u}(t) : [0, T] \to \mathbb{R}^d$ such that the unique solution to Equation (6.1) starting at $\boldsymbol{x}(0) = \boldsymbol{s}$ satisfies $\boldsymbol{x}(T) = \boldsymbol{t}$.

We will show that this problem is decidable in polynomial time, by reduction to the Continuous Orbit Problem. The *Continuous Orbit Problem* consists of deciding whether, given initial and target points $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^n$ and a matrix $A \in \mathbb{R}^{n\times n}$, the unique solution $\boldsymbol{x}(t) = \exp(At)\boldsymbol{s}$ of the differential equation

$$\begin{cases} \dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) \\ \boldsymbol{x}(0) = \boldsymbol{s} \end{cases} \tag{6.2}$$

ever hits $\boldsymbol{t}$. This problem was shown to be decidable in polynomial time in [41, 28].

Before proceeding, we prove a few simple definitions and preliminary results.

Lemma 6.5. *The solution to Equation* (6.1) *is given by*

$$\boldsymbol{x}(t) = \exp(At)\left(\boldsymbol{x}(0) + \int_0^t \exp(-Ay)B\boldsymbol{u}(y)\,dy\right).$$

Proof. Let $\boldsymbol{z}(t) = \exp(-At)\boldsymbol{x}(t)$. Then

$$\begin{aligned} \dot{\boldsymbol{z}}(t) &= -A\boldsymbol{z}(t) + \exp(-At)\dot{\boldsymbol{x}}(t) \\ &= -A\exp(At)\boldsymbol{x}(t) + \exp(-At)A\boldsymbol{x}(t) + \exp(-At)B\boldsymbol{u}(t) \\ &= \exp(-At)B\boldsymbol{u}(t) \\ \Rightarrow \boldsymbol{x}(t) &= \exp(At)\boldsymbol{z}(t) = \exp(At)\left(\boldsymbol{x}(0) + \int_0^t \exp(-Ay)B\boldsymbol{u}(y)\,dy\right). \end{aligned}$$

□

Lemma 6.6. *Let $\mathcal{V}$ be a vector subspace of $\mathbb{R}^n$ and $f : \mathbb{R}_0^+ \to \mathcal{V}$ be a measurable function. Then, for any $t \geq 0$, $\int_0^t f(y)dy \in \mathcal{V}$.*

Proof. Let $\boldsymbol{r} \in \mathcal{V}^\perp$. Then

$$\boldsymbol{r}^T \int_0^t f(y)\, dy = \int_0^t \boldsymbol{r}^T f(y)\, dy = \int_0^t 0\, dy = 0$$

and therefore $\int_0^t f(y)dy \in (\mathcal{V}^\perp)^\perp = \mathcal{V}$, because $\mathcal{V}$ is closed. □

The matrix

$$C = \begin{pmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{pmatrix} \tag{6.3}$$

is called the controllability matrix for the LTI system defined in Equation (6.1). We will denote its image by $\Im(C)$.

Lemma 6.7. *The following are equivalent:*

1. $\boldsymbol{r} \in \Im(C)^\perp$.
2. $\boldsymbol{r}^T \exp(-Ay)B$ *is identically zero on* $[0, \infty)$.
3. *There exists* $T > 0$ *for which* $\boldsymbol{r}^T \exp(-Ay)B$ *is identically zero on* $[0, T]$.

Proof. We show that (1) ⇒ (2) and that (3) ⇒ (1). It is trivial that (2) ⇒ (3). By the Cayley-Hamilton Theorem, $\boldsymbol{r} \in \Im(C)^\perp \Rightarrow \boldsymbol{r}^T A^k B = 0$ for any $k \geq 0$. The first implication follows from the power series definition of matrix exponentials. For the second implication, note that if $f(y) \triangleq \boldsymbol{r}^T \exp(-Ay)B$ is identically zero on $[0, T]$ then $0 = f^{(k)}(0) = \boldsymbol{r}^T(-A)^k B$. □

Proposition 6.8. *For any* $T \geq 0$ *and* $\boldsymbol{v} \in \Im(C)$*, there exists a piecewise constant function* $\boldsymbol{u} : [0, T] \to \mathbb{R}^d$ *such that*

$$\int_0^T \exp(-Ay)B\boldsymbol{u}(y)\, dy = \boldsymbol{v}. \tag{6.4}$$

Proof. Let $\mathcal{H}$ denote the space of piecewise constant functions mapping $[0, T]$ to $\mathbb{R}^d$, and consider the function $g : \mathcal{H} \to \mathcal{V}$ defined by

$$g(\boldsymbol{u}) = \int_0^T \exp(-Ay)B\boldsymbol{u}(y)\, dy.$$

It is clear from Lemma 6.6 and from Lemma 6.7 that $\Im(g) \subseteq \Im(C)$. To show the converse containment, let $\boldsymbol{r} \in \Im(g)^\perp$. Then, for all $\boldsymbol{u} : [0, T] \to \mathbb{R}^d$,

$$0 = \boldsymbol{r}^T \int_0^T \exp(-Ay)B\boldsymbol{u}(y)\, dy = \int_0^T \boldsymbol{r}^T \exp(-Ay)B\boldsymbol{u}(y)\, dy$$

and therefore $\boldsymbol{r}^T \exp(-Ay)B$ must by identically zero on $[0, T]$, due to the arbitrarity of $\boldsymbol{u}$. Thus, we can conclude from Lemma 6.7 that $\boldsymbol{r} \in \Im(C)^\perp$, which implies that $\Im(g)^\perp \subseteq \Im(C)^\perp$, due to the arbitrarity of $\boldsymbol{r}$, implying that $\Im(g) = \Im(C)$. □

We can finally prove the main result of this section.

Theorem 6.9. *The continuous point-to-point controllability problem for LTI systems reduces to the continuous orbit problem in polynomial time, and therefore is in* ***PTIME****.*

Proof. Consider the quotient vector space $\mathbb{R}^n/\Im(C)$. Noting that $\Im(C)$ is invariant under A, A induces a well-defined linear operator on $\mathbb{R}^n/\Im(C)$. We claim that $\boldsymbol{t}+\Im(C)$ is in the orbit of $\boldsymbol{s}+\Im(C)$ by A if and only if we can control Equation (6.1) from $\boldsymbol{s}$ to $\boldsymbol{t}$.

If we can control Equation (6.1) from $\boldsymbol{s}$ to $\boldsymbol{t}$ then, due to Lemma 6.5, there exists $T > 0$ such that

$$\boldsymbol{t} = \boldsymbol{x}(T) = \exp(AT)\left(\boldsymbol{s} + \int_0^T \exp(-Ay)B\boldsymbol{u}(y)\,dy\right)$$

for some control function $\boldsymbol{u} : [0, T] \to \mathbb{R}^d$, so

$$\boldsymbol{t} + \Im(C) = \exp(AT)(\boldsymbol{s} + \Im(C)),$$

as

$$\int_0^T \exp(-Ay)B\boldsymbol{u}(y)\,dy \in \Im(C)$$

due to Lemma 6.6.

On the other hand, suppose that $\boldsymbol{t}+\Im(C)$ is in the orbit of $\boldsymbol{s}+\Im(C)$ by A. Then, there exist $T > 0$ and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \Im(C)$ such that

$$\boldsymbol{t} + \boldsymbol{v}_2 = \exp(At)\left(\boldsymbol{s} + \boldsymbol{v}_1\right).$$

Due to Equation (6.4), there exists a control function $\boldsymbol{u} : [0, T] \to \mathbb{R}^d$ such that

$$\int_0^T \exp(-Ay)B\boldsymbol{u}(y)\,dy = \boldsymbol{v}_1 - \exp(-AT)\boldsymbol{v}_2$$

and so

$$\begin{aligned}\boldsymbol{x}(T) &= \exp(AT)\left(\boldsymbol{s} + \int_0^T \exp(-Ay)B\boldsymbol{u}(y)\,dy\right)\\ &= \exp(AT)\left(\boldsymbol{s} + \boldsymbol{v}_1 - \exp(-AT)\boldsymbol{v}_2\right)\\ &= \exp(AT)\left(\boldsymbol{s} + \boldsymbol{v}_1\right) - \boldsymbol{v}_2 = \boldsymbol{t}.\end{aligned}$$

□

6.4 Conclusion

Complete state controllability for LTI systems has long been characterised by Kalman's criterion, which states that such a system is controllable (that is, any state can be controlled to any other one in finite time) if and only if the controllability matrix has full row rank. We studied the hardness of deciding controllability for a given pair of states. In particular, if the set of controls is rich enough (finite union of convex polytopes), we showed that this problem is undecidable by encoding Hilbert's Tenth Problem. Even when the set of controls is a convex polytope, we proved Skolem-hardness (actually, we proved Positivity-hardness, which is stronger). The problem becomes decidable when the set of controls is a linear subspace, by reduction to the Continuous Orbit Problem.

It would be interesting to get a tighter characterisation of the hardness of this problem when the set of controls is a convex polytope, either by showing undecidability, or by giving an algorithm that queries an oracle to a more famous open problem, such as the Positivity Problem.

Bibliography

[1] I. Adler and P. A. Beling. Polynomial algorithms for linear programming over the algebraic numbers. *Algorithmica*, 12(6):436–457, 1994.

[2] S. Akshay, T. Antonopoulos, J. Ouaknine, and J. Worrell. Reachability problems for Markov chains. *Information Processing Letters*, 115(2):155–158, 2015.

[3] R. Alur. *Principles of Cyber-Physical Systems*. MIT Press, 2015.

[4] L. Babai, R. Beals, J.-Y. Cai, G. Ivanyos, and E. M. Luks. Multiplicative equations over commuting matrices. In *SODA*, pages 498–507. ACM/SIAM, 1996.

[5] A. Bacciotti and L. Mazzi. Stability of dynamical polysystems via families of Lyapunov functions. *Nonlinear Analysis*, 67:2167–2179, 2007.

[6] A. Baker. *Transcendental Number Theory*. Cambridge University Press, 1975.

[7] A. Baker and G. Wüstholz. Logarithmic forms and group varieties. *Journal für die reine und angewandte Mathematik*, 442:19–62, 1993.

[8] S. Basu, R. Pollack, and M.-F. Roy. On the combinatorial and algebraic complexity of quantifier elimination. *Journal of the ACM*, 43(6):1002–1045, 1996.

[9] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in Real Algebraic Geometry*. Springer, 2nd edition, 2006.

[10] J. P. Bell and S. Gerhold. On the positivity set of a linear recurrence. *Israel Journal of Mathematics*, 157:333–345, 2007.

[11] P. Bell, V. Halava, T. Harju, J. Karhumäki, and I. Potapov. Matrix equations and Hilbert's Tenth Problem. *International Journal of Algebra and Computation*, 18(8):1231–1241, 2008.

[12] P. C. Bell, J.-C. Delvenne, R. M. Jungers, and V. D. Blondel. The continuous Skolem-Pisot problem. *Theoretical Computer Science*, 411(40–42):3625–3634, 2010.

[13] A. M. Ben-Amram, S. Genaim, and A. N. Masud. On the termination of integer loops. *ACM Transactions on Programming Languages and Systems*, 34(4):16:1–16:24, 2012.

[14] A.M. Ben-Amram and S. Genaim. On the linear ranking problem for integer linear-constraint loops. In *POPL*, pages 51–62, 2013.

[15] J. Berstel and M. Mignotte. Deux propriétés décidables des suites récurrentes linéaires. *Le Bulletin de la Société Mathématique de France*, 104:175–184, 1976.

[16] V. Blondel and J. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

[17] O. Bournez, M. L. Campagnolo, D. S. Graça, and E. Hainry. Polynomial differential equations compute all real computable functions on computable compact intervals. *Journal of Complexity*, 23(3):317–335, 2007.

[18] O. Bournez, D. S. Graça, and A. Pouly. Computing with polynomial ordinary differential equations. *Journal of Complexity*, 36:106–140, 2016.

[19] A. R. Bradley, Z. Manna, and H.B. Sipma. Termination analysis of integer linear loops. In *CONCUR*, volume 3653 of *LNCS*, pages 488–502. Springer, 2005.

[20] M. Braverman. Termination of integer linear programs. In *CAV*, volume 4144 of *LNCS*, pages 372–385. Springer, 2006.

[21] D. Bridges and P. Schuster. A simple constructive proof of Kronecker's density theorem. *Elemente der Mathematik*, 61:152–154, 2006.

[22] J.-Y. Cai. Computing Jordan normal forms exactly for commuting matrices in polynomial time. *International Journal of Foundations of Computer Science*, 5(3,4):293–302, 1994.

[23] J.-Y. Cai, R. J. Lipton, and Y. Zalcstein. The complexity of the A B C problem. *SIAM Journal on Computing*, 29(6):1878–1888, 2000.

[24] J. Canny. Some algebraic and geometric computations in PSPACE. In *STOC*, pages 460–467. ACM, 1988.

[25] J. W. S. Cassels. *An introduction to Diophantine approximation*. Cambridge University Press, 1965.

[26] E. B. Castelan and J.-C. Hennet. On invariant polyhedra of continuous-time linear systems. *IEEE Transactions on Automatic Control*, 38(11):1680–85, 1993.

[27] H. Y. Chen, S. Flur, and S. Mukhopadhyay. Termination proofs for linear simple loops. In *SAS*, volume 7460 of *LNCS*, pages 422–438. Springer, 2012.

[28] T. Chen, N. Yu, and T. Han. Continuous-time orbit problems are decidable in polynomial-time. *Information Processing Letters*, 115(1):11–14, 2015.

[29] C. Choffrut and J. Karhumäki. Some decision problems on integer matrices. *Theoretical Informatics and Applications*, 39(1):125–131, 2005.

[30] V. Chonev. *Reachability Problems for Linear Dynamical Systems*. PhD thesis, University of Oxford, 2015.

[31] V. Chonev, J. Ouaknine, and J. Worrell. The orbit problem in higher dimensions. In *STOC*, pages 941–950. ACM, 2013.

[32] V. Chonev, J. Ouaknine, and J. Worrell. The polyhedron-hitting problem. In *SODA*, pages 940–956. SIAM, 2015.

[33] V. Chonev, J. Ouaknine, and J. Worrell. On recurrent reachability for continuous linear dynamical systems. In *LICS*, pages 515–524. ACM, 2016.

[34] V. Chonev, J. Ouaknine, and J. Worrell. On the Skolem problem for continuous linear dynamical systems. In *ICALP*, volume 55, pages 100:1–100:13, 2016.

[35] H. Cohen. *A Course in Computational Algebraic Number Theory*. Springer-Verlag, 1993.

[36] M. Colón and H. Sipma. Synthesis of linear ranking functions. In *TACAS*, volume 2031 of *LNCS*, pages 67–81. Springer, 2001.

[37] B. Cook, A. Podelski, and A. Rybalchenko. Termination proofs for systems code. In *PLDI*, pages 415–426. ACM, 2006.

[38] W. J. Culver. On the existence and uniqueness of the real logarithm of a matrix. *Proceedings of the American Mathematical Society*, 17:1146–1151, 1966.

[39] G. Everest, A. J. van der Poorten, I. Shparlinski, and T. Ward. *Recurrence Sequences*. American Mathematical Society, 2003.

[40] J.-H. Evertse. On sums of S-units and linear recurrences. *Compositio Mathematica*, 53(2):225–244, 1984.

[41] E. Hainry. Reachability in linear dynamical systems. In *CiE*, volume 5028 of *LNCS*, pages 241–250. Springer, 2008.

[42] V. Halava. Decidable and undecidable problems in matrix theory. Technical Report 127, Turku Centre for Computer Science, 1997.

[43] V. Halava, T. Harju, and M. Hirvensalo. Positivity of second order linear recurrent sequences. *Discrete Applied Mathematics*, 154(3):447–451, 2006.

[44] V. Halava, T. Harju, M. Hirvensalo, and J. Karhumäki. Skolem's problem – on the border between decidability and undecidability. Technical Report 683, Turku Centre for Computer Science, 2005.

[45] G. Hansel. A simple proof of the Skolem-Mahler-Lech theorem. In *ICALP*, volume 194 of *LNCS*, pages 244–249. Springer, 1985.

[46] M. A. Harrison. *Introduction to Formal Language Theory*. Addison-Wesley, 1st edition, 1978.

[47] B. W. Helton. Logarithms of matrices. *Proceedings of the American Mathematical Society*, 19:733–738, 1968.

[48] T. A. Henzinger. The theory of hybrid automata. In *LICS*, pages 278–292. IEEE Computer Society, 1996.

[49] T. A. Henzinger, P. W. Kopke, A. Puri, and P. Varaiya. What's decidable about hybrid automata? In *STOC*, pages 373–382. ACM, 1995.

[50] R. Kannan and R. J. Lipton. Polynomial-time algorithm for the orbit problem. *Journal of the ACM*, 33(4):808–821, 1986.

[51] L. Khachiyan and L. Porkolab. Integer optimization on convex semialgebraic sets. *Discrete Computational Geometry*, 23:207–224, 2000.

[52] S. A. Kurtz and J. Simon. The undecidability of the generalized Collatz problem. In *TAMC*, volume 4484 of *LNCS*, pages 542–553. Springer, 2007.

[53] V. Laohakosol and P. Tangsupphathawat. Positivity of third order linear recurrence sequences. *Discrete Applied Mathematics*, 157(15):3239–3248, 2009.

[54] C. Lech. A note on recurring series. *Arkiv för Matematik*, 2:417–421, 1953.

[55] A. K. Lenstra, H. W. Lenstra Jr., and László Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen*, 261:515–534, 1982.

[56] L. Liu. Positivity of three-term recurrence sequences. *The Elecronic Journal of Combinatorics*, 17(1), 2010.

[57] A. Macintyre and A. J. Wilkie. On the decidability of the real exponential field. In *Kreiseliana. About and Around Georg Kreisel*, pages 441–467. A K Peters, 1996.

[58] K. Mahler. Eine arithmetische Eigenschaft der Taylor Koeffizienten rationaler Funktionen. *Proceedings of the Royal Netherlands Academy of Arts and Sciences*, 38:50–60, 1935.

[59] G. Malajovich. An effective version of Kronecker's theorem on simultaneous diophantine approximation. Technical report, Universidade Federal do Rio de Janeiro, 1996.

[60] A. Markov. On certain insoluble problems concerning matrices. *Proceedings of the USSR Academy of Sciences*, 57(6):539–542, 1947.

[61] D. W. Masser. Linear relations on algebraic groups. In *New Advances in Transcendence Theory*. Cambridge University Press, 1988.

[62] Y. Matiyasevich. *Hilbert's 10th Problem*. MIT Press, 1993.

[63] K. A. Mikhailova. The occurrence problem for direct products of groups. *Matematicheskii Sbornik*, 70(112):241–251, 1966.

[64] M. Newman. Two classical theorems on commuting matrices. *Journal of Research of the National Bureau of Standards*, 71 B(2, 3):69–71, 1967.

[65] J. Ouaknine, A. Pouly, J. Sousa-Pinto, and J. Worrell. Solvability of matrix-exponential equations. In *LICS*, pages 798–806. ACM, 2016.

[66] J. Ouaknine, J. Sousa-Pinto, and J. Worrell. On termination of integer linear loops. In *SODA*, pages 957–969. SIAM, 2015.

[67] J. Ouaknine, J. Sousa-Pinto, and J. Worrell. On the polytope escape problem for continuous linear dynamical systems. In *HSCC*, pages 11–17. ACM, 2017.

[68] J. Ouaknine and J. Worrell. On the positivity problem for simple linear recurrence sequences. In *ICALP*, volume 8573 of *LNCS*, pages 318–329. Springer, 2014.

[69] J. Ouaknine and J. Worrell. Positivity problems for low-order linear recurrence sequences. In *SODA*, pages 366–379. SIAM, 2014.

[70] J. Ouaknine and J. Worrell. Ultimate positivity is decidable for simple linear recurrence sequences. In *ICALP*, volume 8573 of *LNCS*, pages 330–341. Springer, 2014.

[71] V. Pan. Optimal and nearly optimal algorithms for approximating polynomial zeros. *Computers & Mathematics with Applications*, 31(12):97–138, 1996.

[72] M. S. Paterson. Undecidability in 3×3 matrices. *Journal of Mathematics and Physics*, 49(1):105–107, 1970.

[73] A. Podelski and A. Rybalchenko. Transition invariants. In *LICS*, pages 32–41, 2004.

[74] I. Potapov and P. Semukhin. Vector reachability problem in $SL(2, \mathbb{Z})$. In *MFCS*, volume 58, pages 84:1–84:14, 2016.

[75] I. Potapov and P. Semukhin. Decidability of the membership problem for 2×2 integer matrices. In *SODA*, pages 170–186. SIAM, 2017.

[76] R. Rebiha, N. Matringe, and A. V. Moura. Generating asymptotically non-terminant initial variable values for linear diagonalizable programs. In *SCSS*, pages 81–92, 2013.

[77] J. Renegar. On the computational complexity and geometry of the first-order theory of the reals. *Journal of Symbolic Computation*, 13(3):255–352, 1992.

[78] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois Journal of Mathematics*, 6(1):64–94, 1962.

[79] G. Rozenberg and A. Salomaa. *Cornerstones of Undecidability*. Prentice Hall, 1994.

[80] A. Salomaa. Growth functions of Lindenmayer systems: Some new approaches. In *Automata, Languages, Development*. North-Holland, 1976.

[81] S. Sankaranarayanan, T. Dang, and F. Ivancic. A policy iteration technique for time elapse over template polyhedra. In *HSCC*, volume 4981 of *LNCS*, pages 654–657. Springer, 2008.

[82] W. J. Savitch. Relationships between nondeterministic and deterministic tape complexities. *Journal of Computer and System Sciences*, 4:177–192, 1970.

[83] H. P. Schlickewei. The $\mathfrak{p}$-adic Thue-Siegel-Roth-Schmidt Theorem. *Archiv der Mathematik*, 29:267–270, 1977.

[84] C. Shannon. Mathematical Theory of the Differential Analyzer. *Journal of Mathematics and Physics*, 20:337–354, 1941.

[85] T. Skolem. Ein Verfahren zur Behandlung gewisser exponentialer Gleichungen. In *Scandinavian Congress of Mathematicians*, pages 163–188, 1934.

[86] I. Stewart and D. Tall. *Algebraic Number Theory and Fermat's Last Theorem*. A K Peters, 2002.

[87] A. Tarski. *A Decision Method for Elementary Algebra and Geometry*. University of California Press, 1951.

[88] A. Tiwari. Termination of linear programs. In *CAV*, volume 3114 of *LNCS*, pages 70–82. Springer, 2004.

[89] A. J. van der Poorten and H. P. Schlickewei. The growth conditions for recurrence sequences. *Macquarie Mathematics Report*, (82-0041), 1982.

[90] N. K. Vereshchagin. The problem of appearance of a zero in a linear recurrence sequence. *Mathematical Notes*, 38(2):177–189, 1985.

[91] E. Wermuth. Two remarks on matrix exponentials. *Linear Algebra and its Applications*, 117:127–132, 1989.