Reachability and Escape Problems in Linear Dynamical Systems

Julian Reuben D'Costa
Keble College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2024

To my family, who have cheered me on throughout this journey.

Acknowledgements

I am deeply grateful to my advisors, Ben Worrell, Joël Ouaknine, and Stefan Kiefer for their guidance and support throughout my DPhil. I first met Ben and Joël through a chain of lucky coincidences five years ago and I have delighted in working with them ever since. It has been an honour to learn from them, and I am grateful for their patience, encouragement, and the many opportunities they have given me. I was not fortunate enough to also work directly with Stefan, but I am grateful for his helpfulness and care in his role as my official co-supervisor.

I would like to thank Jussi Westergren for many stimulating conversations and the support that made it possible for me to study at Oxford.

Most of my research progress has been during or directly inspired by discussions with my wonderful collaborators Toghrul, Mahmoud, Eike, Engel, Sadegh, Rupak, and of course Ben and Joel.

I am grateful to my examiners Alessandro Abate (Transfer and Confirmation), Igor Potapov (Viva) and Christoph Haase (Transfer, Confirmation and Viva) for their thoughtful comments, which have significantly improved this thesis.

I would also like to thank students who have come before me, in particular Ventsi, Joao, Ines, and Mehran for some very valuable advice and their theses, which I have learned a great deal from.

Finally, I would like to thank all my friends, especially Emanuil, Mindy, Hugo, Paavni, Duncan, Andrew, Carlota, Roddy, Quratulain, Ferdie, Sarah, Jess D., Bob, Jess B., Satwik, Abheek, Raj, Rohith, Gaurang, Vasanth, Sabyasachi, Nishit, and Adit for making these four years a time of exploration, growth and joy.

Abstract

In this thesis we explore decision problems for linear dynamical systems, which are a fundamental model in theoretical computer science, software verification and control theory. We focus on the reachability and escape problems in discrete and continuous linear dynamical systems, and their generalizations to pseudo-orbits, which are a model of computation with limited precision.

We consider the problem of deciding whether a finitely generated matrix semigroup contains a non-negative matrix, and show that it is decidable subject to Schanuel's conjecture when the matrices commute, and undecidable in general. We show unconditional decidability of eventual non-negativity for a single matrix, solving a problem posed by S. Akshay [2].

In the setting of discrete pseudo-orbits, we show that the analogues of the classical orbit and Skolem problems are decidable. For diagonalisable linear dynamical systems, we show that the pseudo-reachability problem is decidable for arbitrary semialgebraic targets. We show that the similar robust reachability problem is decidable in full generality. For continuous pseudo-orbits, we show the same results assuming Schanuel's conjecture.

Finally, we establish a uniform upper bound on the number of iterations it takes for every orbit of a matrix to escape a compact semialgebraic set. For a rational matrix and set defined over rational data, the bound is doubly exponential in the dimension, and singly exponential in the rest of the data. We show that this bound is tight by providing a matching lower bound.

Statement of Originality

The contents of this thesis are based on my publications at the 46th and 47th International Symposia on Mathematical Foundations of Computer Science, *MFCS 2021* [28] and *2022* [29, 31] and the 41st International Symposium on Theoretical Aspects of Computer Science, *STACS 2024* [34].

Contents

Chapter 1

Introduction

A *discrete-time linear dynamical system* (LDS) is given by an update matrix $A \in \mathbb{Q}^{d\times d}$ and a starting point s in $\mathbb{Q}^d$. The system evolves according to the dynamics $x \mapsto Ax$. The *orbit* or *trajectory* of the LDS is the sequence of points $s, As, A^2s, \ldots$. Linear dynamical systems are a fundamental model in control theory, verification of programs and cyber-physical systems. In this thesis we study two kinds of problems for LDS: reachability and escape.

The reachability problem for LDS asks whether a given target set T is reachable from a given initial set S. That is, given a matrix A, a target set T, and an initial set S, the reachability problem asks whether there exists *some* s in S such that the orbit of the LDS defined by (A, s) intersects T. Non-reachability is often proved by exhibiting an invariant set that contains the initial set and is disjoint from the target set. An invariant set is a set I that is preserved by the dynamics, i.e. $AI \subseteq I$. All trajectories starting in I will remain trapped in I forever. The dual of an invariant set is an *escape set*, which is an initial set which is left by *every* trajectory. An *Escape Problem* asks whether a given initial set S is an escape set for the matrix A, i.e., whether for *all* s in S there exists a time n such that $A^n s$ is in the complement of S.

In the context of verification of a program or a cyber-physical system, we are usually interested in proving *safety* and *liveness* properties of the system. For a dynamical system, we may interpret a “property” as delineating a subset of the set of all possible trajectories. Safety properties are those that should always be true of the infinite trajectory, such as the absence of collisions in a traffic system. Liveness properties are those that should eventually hold, such as the arrival of a train at a station. It can be shown that every property can be expressed as a conjunction of safety and liveness properties [3]. The problems of reachability and escape we consider for LDS are highly relevant to proving safety and liveness properties respectively.

When modelling a system as an LDS, we may express a safety property as non-reachability of a "bad" set. Conversely, we may express a liveness property as eventual escape from an initial set.

One may consider similar problems for models more expressive than LDS, but even slight increases in power lead to Turing-complete models of computation. For example, allowing quadratic updates leads to undecidability of the reachability problem [44] and allowing piecewise linear updates leads to undecidability of the loop termination problem (an instance of an escape problem) [11].

Overview of reachability results for LDS. Perhaps the simplest possible reachability problem on LDS is the *orbit problem*, which asks whether a given point t is reachable from an initial point s [49]. In 1986, Kannan and Lipton [53] showed that the orbit problem can be decided in polynomial time.

A natural generalization of the orbit problem, the *Skolem problem*[1] in which we ask whether a hyperplane is reachable from an initial point, turns out to be notoriously difficult and remains open after many decades [90, 77]. A breakthrough occurred in the mid-1980s, when Mignotte *et al.* [69] and Vereshchagin [94] independently showed decidability in dimension 4 or less. More recently, Bilu et al [14] showed decidability for the Skolem problem with *diagonalisable* matrices, assuming the p-adic Schanuel conjecture and the Skolem Conjecture (aka Exponential Local-Global Principle) in order to guarantee termination of their algorithm. This (conditionally) also solves the last open case of the Skolem problem for dimension 5.

Similar to the Skolem problem is the Positivity problem, which asks whether a halfspace is reachable from an initial point. In 2014, Ouaknine and Worrell [74] showed that solving the Positivity problem in dimension 6 or higher would require a significant breakthrough in Diophantine approximation. The reachability problem for arbitrary *semialgebraic* initial and target sets S and T can be reduced to querying whether an initial point reaches an intersection of halfspaces [56].

While progress on reachability has been bottlenecked by the Skolem and Positivity problems in light of the result above, no undecidability results are known. In the setting of *multiple reachability*, however, the following problem was shown to be undecidable: given an update matrix A, a semialgebraic starting set S, a positive integer k, and a target hyperplane H, decide if there exists s in S such that the orbit $(A^n s)_{n\in\mathbb{N}}$ reaches H at least k times [55]. The undecidability proof is based on a reduction from Hilbert's 10th problem.

[1]The Skolem problem is traditionally formulated in terms of linear recurrence sequences, but we will show in the next chapter that it is Turing equivalent to the hyperplane reachability problem

Overview of escape results for LDS. The escape problem has been extensively studied in the context of universal termination of affine loop programs. This corresponds to the following escape problem: given a matrix A and a *polytope* S, decide whether S is an escape set for A. This problem was solved considering all real points in S by Tiwari in 2004 [92], for all rational points in S by Braverman in 2006 [17], and for all integer points in S by Hosseini et al [51] in 2019. For semialgebraic S, the escape problem remains open, although [71] showed decidability for the case where S is *compact* via two unbounded searches. Building on this, [30] characterized the complexity of the escape problem for compact semialgebraic sets by showing that it is a complete problem for a certain fragment of the theory of the reals with two quantifiers.

Overview of reachability problems for matrix semigroups. Linear dynamical systems involve repeated applications of a single matrix. A more general setting is that of *finitely generated matrix semigroups*, where we consider the semigroup generated by a set of matrices $A_1, \ldots, A_k$. A classical reachability problem in this setting is reachability of a given matrix, known as the *Membership Problem*.

The Membership Problem for finitely generated matrix semigroups asks, given square matrices $A, A_1, \ldots, A_k$ of the same dimension and with rational entries, whether A lies in the semigroup generated by $A_1, \ldots, A_k$. The problem was shown to be undecidable by Markov in the 1940s [64], thereby becoming one of the first instances of a natural undecidable mathematical problem. The problem is decidable under the assumption that the matrices $A_1, \ldots, A_k$ commute [4].

In between the commuting and non-commuting cases lies the problem of *Solvability of Matrix Equations*. Here we fix the order of the matrices, ie ask whether there exist integers $n_1, \ldots, n_k$ such that $A = A_1^{n_1} \cdots A_k^{n_k}$. For commuting matrices this is equivalent to Membership. For two non-commuting matrices, the problem is decidable. In general, the problem is undecidable via a reduction to Hilbert's tenth problem [7].

There are many variants of the Membership Problem. In the *Mortality Problem* one asks whether the zero matrix lies in a finitely generated matrix semigroup. This problem is undecidable already for 3×3 matrices [80]. In a different direction, for three matrices in fixed order, Mortality is equivalent to Skolem in any dimension [10].

Meanwhile, the *Identity Problem* asks to determine whether the identity matrix lies in a given finitely generated matrix semigroup. The latter problem is undecidable in general but decidable for certain nilpotent and low-order matrix groups [8, 36, 37, 57].

Analogous to the orbit, Skolem and Positivity problems for LDS are the *Vector Reachability*, the *Hyperplane Reachability (aka Scalar Reachability)* and the *Halfspace Reachability* problems for matrix semigroups. Here we have the set of matrices $\{A_1, \ldots, A_k\}$, two vectors u and v and (possibly) a scalar λ, and we ask whether there exists a matrix A in the semigroup generated by $A_1, \ldots, A_k$ such that $Av = u$, $u^\top Av = \lambda$ or $u^\top Av \geq \lambda$ respectively. All these problems are undecidable in general, with some known decidability results for low dimensions and special cases [9, 47].

We mention a final problem that combines matrix semigroups and escape: the Linear Subspace Escape Problem. Given a linear subspace U of $\mathbb{Q}^d$, and a set of matrices $A_1, \ldots, A_k$, is U an escape set for the semigroup? Formally, we ask (dually) whether there exists a *trapped sequence*: an initial point s and an infinite sequence of matrices $A_{i_1}, A_{i_2}, \ldots$ from the set such that for every $k \in \mathbb{N}$, $A_{i_k} \cdots A_{i_1} s$ is in U. This problem (in its dual form) was shown to be decidable under the name of the *Invariance Problem* [38].

1.1 Contributions of this thesis

Given the known barriers to further progress on the classical reachability problem, we consider two variants: First, in chapter 3 we study matrix reachability problems, investigating reachability of the orthant of *non-negative matrices*. Specifically, we consider the semigroup generated by a set of matrices $A_1, \ldots, A_k$ and ask whether it contains a non-negative matrix. We show that if they commute, this problem is decidable subject to Schanuel's conjecture. If they don't commute, the problem is undecidable. For a single matrix M, we show unconditional decidability, which is in contrast to the notorious fact that it is not known how to determine, for any specific matrix index (i, j), whether the sequence $(M^n)_{i,j}$ is ultimately nonnegative. The latter is equivalent to the Ultimate Positivity problem for linear recurrence sequences, which, like Positivity, is a longstanding open problem.

Second, we consider an inexact version of reachability, where we allow for arbitrarily small amounts of noise in the dynamics. This corresponds to the classical notion of *pseudo-orbits*. We show decidability results for pseudo-reaching a point, a hyperplane and a bounded semialgebraic set in chapter 4. In chapter 5 we use a completely different approach based on the first-order theory of the reals with exponentiation to generalise the results of the previous chapter to arbitrary semialgebraic targets for diagonalisable linear dynamical systems. We also show that our method can be used to reduce the continuous-time pseudo-reachability problem to the (classical) time-bounded reachability problem, which is known to be decidable assuming Schanuel's conjecture. We show the same results for robust reachability, which is another inexact generalisation of reachability allowing for imprecision in the initial state.

In the final chapter (chapter 6) we study the escape problem for discrete-time linear dynamical systems over compact semialgebraic sets. We complement the existing decidability result with a bound on the *time* taken to escape. Specifically, we establish a uniform upper bound on the number of iterations it takes for every orbit of a rational matrix to escape a compact semialgebraic set defined over rational data. Our bound is doubly exponential in the ambient dimension, singly exponential in the degrees of the polynomials used to define the semialgebraic set, and singly exponential in the bitsize of the coefficients of these polynomials and the bitsize of the matrix entries. We show that our bound is tight by providing a matching lower bound.

The introductions to each chapter provide more detailed motivation and background for the results described above. The next chapter introduces the mathematical background for this thesis, and proves a useful result bounding the description size increase when transforming a system into a Jordan basis.

Chapter 2

Mathematical Background

In this chapter we describe the mathematical tools used in this thesis, prove a few results of independent interest, and define the terms and problems mentioned in the introduction.

2.1 Notation

The sets of natural numbers, rational numbers, real numbers, and algebraic numbers are denoted by $\mathbb{N}$, $\mathbb{Q}$, $\mathbb{R}$, and $\overline{\mathbb{Q}}$, respectively.

For any column vector $x = [x_1, x_2, \ldots, x_m]^\top \in \mathbb{R}^m$, we use the notations $\|x\|_2 := \sqrt{x^\top x}$ and $\|x\|_\infty := \max_i |x_i|$ to indicate respectively the two norm and infinity norm of x. For any matrix $A = [a_{ij}]_{i,j} \in \mathbb{R}^{m\times m}$, we define $\|A\|_2$ and $\|A\|_\infty$ to indicate respectively the (induced) two norm and infinity norm of A. Unless otherwise specified, $\|A\|$ denotes the two norm of A. Note that $\|Ax\|_2 \leq \|A\|_2 \|x\|_2$ and $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$ for all $x \in \mathbb{R}^m$. We denote by $\rho(A)$ the spectral radius of a matrix A, which is the largest absolute value of the eigenvalues of A.

We write $\mathbf{0} \in \mathbb{R}^m$ for the zero vector and $\mathbf{1} \in \mathbb{R}^m$ for the all-ones vector. We write $B(c, r)$ for the closed ball (under appropriate norm, usually ℓ_2 or ℓ_∞) of radius r centred around $c \in \mathbb{R}^m$. We denote by $\mathbb{T} \subseteq \mathbb{C}$ the unit circle in the complex plane.

2.1.1 Asymptotic notation

We use the standard asymptotic notation $O(\cdot)$ to describe the growth of functions, especially in the exponent, which we define below. A statement of the form $f(n, d, \tau) = 2^{(d\tau)^{n^{O(1)}}}$ means that there exist positive constants $C_0, C_e, n_0, d_0, \tau_0$ such that for all $n \geq n_0$, $d \geq d_0$, and $\tau \geq \tau_0$, we have $|f(n, d, \tau)| \leq C_0 \cdot 2^{(d\tau)^{n^{C_e}}}$. We use the notation

$O^*(\cdot)$ to hide polylogarithmic factors, i.e. $O^*(n) = O(n(\log n)^C)$ for some constant C.

All logarithms are base 2 unless specified otherwise. Integer inputs to algorithm are specified in binary. Rational inputs are specified as a pair of integers. For a rational number or polynomial represented as a list of integer coefficients, we will refer to the total number of bits required to store these coefficients in binary as the *bitsize*. Note that the bitsize of an integer n is $O(\log n)$.

2.2 Algebraic Numbers

An algebraic number is a complex number that is a root of a non-zero polynomial with integer coefficients. There exists such a polynomial of minimal degree for each algebraic number, called its minimal polynomial, which is irreducible and thus squarefree. The *degree* of an algebraic number is the degree of its minimal polynomial. The *height* of an algebraic number is the maximum of the absolute values of the coefficients of its minimal polynomial.

We will assume the standard representation (see [22]) of an algebraic number by storing the coefficients of its minimal polynomial, and a rational approximation accurate enough to distinguish it from its conjugates - the other roots of its minimal polynomial.

A classical theorem due to Mignotte [68] lets us bound the precision required as polynomial in the degree and bitsize of the coefficients of the minimal polynomial.

Theorem 2.1 (Mignotte)**.** *Let $p \in \mathbb{Z}[x]$ be a square-free univariate polynomial of degree at most d, whose coefficients have absolute value bounded by H. Let $\alpha \neq \beta$ be distinct roots of p. Then*

$$|\alpha - \beta| > \frac{\sqrt{3}}{(d+1)^{d+2} H^{d-1}}.$$

Corollary 2.2 ([84, Theorem 4])**.** *Let $p \in \mathbb{Q}[x]$ be a univariate polynomial of degree d and bitsize τ. Let $\alpha \neq \beta$ be distinct roots of p. Then*

$$|\alpha - \beta| > \frac{1}{2^{(d\tau)^{O(1)}}}.$$

The proof essentially involves factorizing the polynomial into irreducible factors and removing duplicate factors to obtain a square-free polynomial.

With this representation we can perform arithmetic operations and test equality on algebraic numbers in time polynomial in their representation (see, e.g., [22] [78] [59]).

We will need to bound the modulus of an algebraic number away from 1. This is achieved by combining the Mignotte bound with a bound on the height of the resultant of two polynomials, proved in [15, Theorem 10].

Lemma 2.3. *Let λ be a complex algebraic number of degree at most d and height at most 2^τ. Assume that $|\lambda| \neq 1$. Then we have $||\lambda| - 1| > 2^{-(\tau d)^{O(1)}}$.*

We prove the following more explicit version of Lemma 2.3:

Lemma 2.4. *Let λ be a complex algebraic number of degree d and height H. Assume that $|\lambda| \neq 1$. Then*

$$||\lambda| - 1| > \frac{1}{\sqrt{3}(2d+2)^{2d+3}2^{2d}(d!)^{2d}(d+1)^{2d(d-1)}H^{2d^2}}.$$

Proof. The numbers λ and $\bar{\lambda}$ are roots of the same minimal polynomial P of degree d and height H. It follows that the number $|\lambda|^2 = \lambda\bar{\lambda}$ is a root of the polynomial $Q(x) = \mathrm{res}_z\left(P(x), x^d P(z/x)\right)$, where $\mathrm{res}_z(A, B)$ denotes the resultant of the polynomials $A, B \in \mathbb{Q}[x][z]$ with coefficients in the integral domain $\mathbb{Q}[x]$ (see [22, p. 159]).

The degree of $Q(x)$ is at most $2d$. By [15, Theorem 10] the height of $Q(x)$ is bounded by

$$H' = d!(d+1)^{d-1}H^d.$$

The polynomial $Q(x)(x-1)$ has degree at most $2d+1$ and height at most $2H'$.

It follows from Theorem 2.1 that

$$\left||\lambda|^2 - 1\right| > \frac{\sqrt{3}}{(2d+2)^{2d+3}(2H')^{2d}}.$$

Note that $||\lambda|^2 - 1| = ||\lambda| - 1| \cdot ||\lambda| + 1|$. If $|\lambda| > 2$ then the claim is trivial, so we may assume that $||\lambda| + 1| \leq 3$, yielding

$$||\lambda| - 1| > \frac{\sqrt{3}}{3(2d+2)^{2d+3}(2H')^{2d}} = \frac{1}{\sqrt{3}(2d+2)^{2d+3}2^{2d}(d!)^{2d}(d+1)^{2d(d-1)}H^{2d^2}}.$$

□

2.2.1 Groups of Multiplicative Relations

We will also need the following theorem due to Masser [66] which shows that the integer multiplicative relations between given complex algebraic numbers can be elicited in polynomial space. For a proof see [19, 41].

Theorem 2.5 (Masser's theorem). *Let $(\lambda_1, \ldots, \lambda_m)$ be complex algebraic numbers. Consider the free Abelian group under addition*

$$L = \{(n_1, \ldots, n_m) \in \mathbb{Z}^m \mid \lambda_1^{n_1} \cdot \cdots \cdot \lambda_m^{n_m} = 1\}.$$

Then one can compute in polynomial space a basis $(\beta_1, \ldots, \beta_s) \in (\mathbb{Z}^m)^s$ for L. Moreover, the integer entries of the basis elements β_j are bounded polynomially in the representations of the minimal polynomials of $\lambda_1, \ldots, \lambda_m$ and singly exponentially in m.

Note that this is a bound on the absolute values of the entries of the integer lattice, not their bitsizes, which means there are only polynomially many *possibilities* (for fixed m), which allows for an algorithm that uses exhaustive search.

2.3 Linear Algebra

Analysing the powers of a linear map is easiest if it can be written as a diagonal matrix in some basis. Unfortunately, this is not always possible. The Jordan normal form (JNF) is a generalization of diagonalisation that applies to any square matrix, even if it is non-diagonalisable, at the cost of needing to work in the splitting field of the characteristic polynomial. For a rational matrix, we usually work in the field of algebraic numbers.

Jordan Decomposition. For a given rational square matrix A one can compute *change of basis matrix* Q and *Jordan normal form* J so that $A = QJQ^{-1}$ and $J = \mathrm{diag}(J_1, J_2, \cdots, J_z)$ with J_i representing the i^{th} Jordan block taking the following form

$$J_i = \begin{bmatrix} \Lambda_i & 1 & 0 & \ldots & 0 & 0 \\ 0 & \Lambda_i & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \Lambda_i & 1 \\ 0 & 0 & 0 & \ldots & 0 & \Lambda_i \end{bmatrix}, \tag{2.1}$$

where Λ_i denotes the i^{th} eigenvalue of A. The size of J_i is equal to the multiplicity of the eigenvalue Λ_i and is denoted by $\kappa(\Lambda_i)$.

Real Jordan form. For any $A \in \mathbb{R}^{n \times n}$ having complex eigenvalues, matrices Q and J in the Jordan normal form could have complex entries. In this case, the complex eigenvalues form complex conjugate pairs and give a *real Jordan form*: there are *real* matrices Q and J such that $A = QJQ^{-1}$ and $J = \mathrm{diag}(J_1, J_2, \cdots, J_z)$. The matrix J_i represents the i^{th} real Jordan block corresponding to either a real eigenvalue Λ_i or a complex pair $\Lambda_i = a_i \pm jb_i$. It is equal to (2.1) for real Λ_i and has the following form for the complex pair $\Lambda_i = a_i \pm jb_i$,

$$J_i = \begin{bmatrix} \Lambda_i & I_{2\times 2} & 0_{2\times 2} & \dots & 0_{2\times 2} & 0_{2\times 2} \\ 0_{2\times 2} & \Lambda_i & I_{2\times 2} & \dots & 0_{2\times 2} & 0_{2\times 2} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & \dots & \Lambda_i & I_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & \dots & 0_{2\times 2} & \Lambda_i \end{bmatrix}, \tag{2.2}$$

where with abuse of notation, we have indicated $\Lambda_i = \begin{bmatrix} a_i & -b_i \\ b_i & a_i \end{bmatrix}$. $I_{2\times 2}$ and $0_{2\times 2}$ denote identity and fully zero matrices of size 2 by 2.

Computing matrix powers. If $A = QJQ^{-1}$, then we have $A^n = QJ^nQ^{-1}$ for $n \in \mathbb{N}$, where $J^n = \mathrm{diag}(J_1^n, J_2^n, \dots, J_z^n)$ and

$$J_i^n = \begin{bmatrix} \Lambda_i^n & n\Lambda_i^{n-1} & \binom{n}{2}\Lambda_i^{n-1} & \cdots & \binom{n}{k-1}\Lambda_i^{n-k+1} \\ 0 & \Lambda_i^n & n\Lambda_i^{n-1} & \cdots & \binom{n}{k-2}\Lambda_i^{n-k+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & n\Lambda_i^{n-1} \\ 0 & 0 & 0 & \cdots & \Lambda_i^n \end{bmatrix}.$$

Here $k = \kappa(\Lambda_i)$ is the size of the Jordan block J_i.

For a square matrix $A \in \mathbb{R}^{m\times m}$, its corresponding exponential matrix is denoted by e^A and is defined as $e^A := \sum_{k=0}^{\infty} \frac{A^k}{k!}$.

2.3.1 Computing real JNF in polynomial time

Computing the (complex) Jordan normal form of a given rational matrix has historically been seen as a difficult problem, because it is highly numerically unstable. The map from matrices to their Jordan forms is discontinuous - an infinitesimal perturbation that alters one occurrence of a duplicate eigenvalue to create a distinct eigenvalue will change the structure of the Jordan blocks. It is still possible to compute the JNF symbolically representing the eigenvalues in their canonical forms as algebraic numbers. This gives an approximation to the JNF of any desired precision in time polynomial in the digits of precision, the dimension, and the bitsize of the entries of the original matrix (assumed rational). Until recently, the best known method was the algorithm of Cai in 1996 [18] where the polynomial was unspecified but "of degree at least 12 in the dimension" [35].

However there is now a more explicit bound:

Theorem 2.6 (Computing JNF in poly time [35]). *Let A be an $n \times n$ matrix of τ-bit integers. Let b be a desired precision parameter. Then there exists an algorithm that outputs complex matrices J' and V' (with rational real and imaginary parts of entries of bitsize $\tau n^3 + b$) with the following properties.*

- $\|J - J'\| \leq 2^{-b} \|J\|$, $\|V - V'\| \leq 2^{-b} \|V\|$ *for some exact JNF* $A = VJV^{-1}$.
- V' *is relatively well conditioned:* $\|V'\| \cdot \|V'^{-1}\| \leq 2^{O^*(\tau n^3)}$.
- *The algorithm runs in expected* $O^*(n^{\omega+3}\tau + n^4\tau^2 + n^\omega b)$ *bit operations.*

Here the notation O^* denotes suppression of factors polylogarithmic in n, τ and b. The symbol ω is the complexity constant of matrix multiplication, which is known to be less than 3.

We can apply this to a rational matrix of bounded bitsize by factoring out a common denominator (of size at most 2^τ) and increasing precision by τ.

Theorem 2.6 does not specify how the matrix inverse V'^{-1} is computed. However, the matrix V' is well-conditioned, so we can compute V'^{-1} in polynomial time using standard techniques. Alternatively, we can use the algorithm of Cai which essentially uses a "row vector" version of the algorithm for V' to compute the inverse of V' in polynomial time.

We now discuss how to compute the *real* Jordan normal form QJQ^{-1} of A in polynomial time. First compute, in polynomial time, the (complex) Jordan normal form J' and matrices V, V^{-1} such that $A = VJ'V^{-1}$ using the algorithm from [35] or [18].

Computing J: Suppose, without loss of generality, that

$$J' = \operatorname{diag}(J'_1, J'_2, \ldots, J'_{2k-1}, J'_{2k}, J'_{2k+1}, \ldots, J'_{2k+z})$$

where for $1 \leq j \leq k$, the Jordan blocks J_{2j-1} and J_{2j} have the same dimension and have conjugate eigenvalues $\lambda_j = a_j + b_j i$ and $\overline{\lambda} = a_j - b_j i$, respectively. The blocks $J'_{2k+1}, \ldots, J'_{2k+z}$, on the other hand, have real eigenvalues. J is obtained by replacing, for each $1 \leq j \leq k$, $\operatorname{diag}(J'_{2j-1}, J'_{2j})$ with a real Jordan block of the same dimension with $\Lambda = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ and keeping the blocks $J'_{2k+1}, \ldots, J'_{2k+z}$ unchanged.

Computing Q: Let $\kappa(j)$ denote the multiplicity of the Jordan block J'_i for $1 \leq i \leq 2k + z$, and $v_1^1, \ldots, v_{\kappa(1)}^1, \ldots, v_1^{2k}, \ldots, v_{\kappa(2k)}^{2k}, \ldots, v_1^{2k+z}, \ldots, v_{\kappa(2k+z)}^{2k+z} \in \overline{\mathbb{Q}}^m$ be the columns of V. It will be the case that for all $1 \leq j \leq k$ and l, $v_l^{2j-1} = \overline{v_l^{2j}}$ in the sense that $v_l^{2j-1} = x_l^j + y_l^j i$ and $v_l^{2j} = x_l^j - y_l^j i$ for vectors $x_l^j, y_l^j \in \mathbb{R}^m$. Moreover, for

$j > 2k$, $v_l^{2j} \in \mathbb{R}^m$. Finally, columns of Q are obtained from columns of V as follows. For $1 \leq j \leq k$ and all l, replace v_l^{2j-1} with x_l^j and v_l^{2j} with y_l^j and keep v_l^{2k+z} for all l and $m > 0$ unchanged, in the same way the proof of existence of real Jordan normal form proceeds.

Computing Q^{-1}: Summarizing the construction above, Q is obtained from V by replacing columns $x + yi$ and $x - yi$, $x, y \in \mathbb{R}^m$ by x and y, respectively. Since $x = \frac{1}{2}(x+yi)+\frac{1}{2}(x-yi)$ and $y = -\frac{1}{2}i(x+yi)+\frac{1}{2}i(x-yi)$, this construction is linear and we can write $Q = V \cdots T$ for some $T \in \mathbb{C}^{m\times m}$ with entries in $\{\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}i, -\frac{1}{2}i, 1, 0\}$. Moreover, the linear transformation T is clearly invertible: $x + yi = 1 \cdot x + iy$ and $x - yi = 1 \cdot x - (-i)y$, and hence $T^{-1} \in \mathbb{C}^{m\times m}$ with entries in $\{1, i, -i, 0\}$. Finally, compute Q^{-1} via $Q = VT \implies Q^{-1} = T^{-1}V^{-1}$, observing that we already know how to compute V^{-1} in polynomial time.

2.4 Logic and Decidability

2.4.1 First-order logic

A first-order language $\mathcal{L}$ is specified by a set F of function symbols, a set R of relation symbols, and a set C of constant symbols. We also assume an infinite set of variable symbols. A term in the language $\mathcal{L}$ is a well-formed expression built from the variable symbols and the symbols in F and C. Atomic formulas in $\mathcal{L}$ are of the form $t_1 = t_2$ where t_1, t_2 are terms or $r(t_1, \ldots, t_k)$, where $r \in R$ is a relation symbol with arity k, and $t_1, \ldots, t_k$ are terms. Well-formed formulas in $\mathcal{L}$ are constructed from atomic formulas, quantifiers and logical connectives in the usual way. A *sentence* is a formula that does not contain free variables. The *language of rings* $\mathcal{L}_r = \{+, -, \cdot, 0, 1\}$, where $+, -, \cdot$ are binary function symbols, $=$ is the relation symbol and $0, 1$ are constants. The *language of ordered rings* $\mathcal{L}_{or}$ adds the relation symbol $<$ to $\mathcal{L}_r$. Observe that by iterating $+$, $-$, $\cdot$ and 0, 1 we can define formulae corresponding to all rationals in $\mathcal{L}_{or}$.

A *structure* $\mathcal{M}$ for a language $\mathcal{L}$ is given by a nonempty set M called the *universe, domain,* or *underlying set* of $\mathcal{M}$, a function $f^{\mathcal{M}} : M^{n_f} \to M$ for each $f \in F$ (of arity n_f), a set $r^{\mathcal{M}} \subseteq M^{n_r}$ for each $r \in R$, and an element $c^{\mathcal{M}} \in M$ for each $c \in C$. These are known as the *interpretations* of the symbols f, r and c.

Let $\mathbb{R}_0$ denote the structure of the real numbers with the standard arithmetic operations and binary relations. The set $\mathbb{R}$ is the universe of the structure $\mathbb{R}_0$. We will be interpreting formulas of $\mathcal{L}_{or}$ in the structure $\mathbb{R}_0$. What it means for a logic to be *first-order* is that it allows quantification over elements of the universe of a structure, but not over more complicated domains such as subsets of the universe or relations.

Let $\mathcal{M}$ be a structure with universe M whose functions, relations and constants respectively match the function, relation and constant symbols of a language $\mathcal{L}$. We say that $S \subseteq M^d$ is *definable in* $\mathcal{M}$ *with parameters from* $X \subseteq M$ if there exist $k \geq 0$, a formula $\phi \in \mathcal{L}$ with $k+d$ free variables, and $c_1, \ldots, c_k \in X$ such that for all $x \in M^d$, $x \in S \iff \phi(c_1, \ldots, c_k, x)$. We say that S is *definable* in $\mathcal{M}$ if it is definable in $\mathcal{M}$ with parameters from M, and $\mathcal{L}$-*definable* if it is definable in $\mathcal{M}$ with parameters from $X = \emptyset$.

By the (first-order) theory of $\mathcal{M}$, denoted by $\mathrm{Th}(\mathcal{M})$, we mean the set of $\mathcal{L}$-sentences ϕ such that ϕ is true in $\mathcal{M}$, written $\mathcal{M} \models \phi$. The theory $\mathrm{Th}(\mathbb{R}_0)$ is known as the first-order theory of real closed fields.

A theory $\mathrm{Th}(\mathcal{M})$ admits *quantifier elimination* if for every $\mathcal{L}$-formula $\phi(x_1, \ldots, x_n)$ there exists a quantifier-free $\psi(x_1, \ldots, x_n)$ such that for all $a_1, \ldots, a_n \in M$, $\mathcal{M} \models \phi(a_1, \ldots, a_n)$ if and only if $\mathcal{M} \models \psi(a_1, \ldots, a_n)$. We say that the formula ψ is equivalent to ϕ modulo $\mathcal{M}$.

The theory $\mathrm{Th}(\mathcal{M})$ is *decidable* if, given an $\mathcal{L}$-sentence ϕ, it is decidable whether $\phi \in \mathrm{Th}(\mathcal{M})$. Tarski [91] showed that the theory of real closed fields admits quantifier elimination and is therefore decidable. The implication is easy to show: given a sentence ϕ in $\mathcal{L}_{or}$, we first compute an equivalent quantifier-free sentence ψ, which will be a Boolean combination of atomic formulas without any variables, containing constants, arithmetic symbols, and relation symbols - and then check its truth, which in $\mathbb{R}_0$ amounts to checking whether an integer is positive or negative. For example, consider the sentence $\exists x. x^2+1 = 0$. As we know from elementary algebra, this reduces to the quantifier free formula corresponding to checking whether the discriminant $b^2 - 4ac$ of this quadratic polynomial is greater than or equal to zero. In this case $0^2 - 4 \cdot 1 \cdot 1 < 0$, so it is not, and the original sentence has been shown to be false.

2.4.2 Decidability of the theory of real exponentiation

We denote by $\mathcal{L}_e$ the language of real exponentiation, which extends $\mathcal{L}_{or}$ with the unary function symbol $\exp(\cdot)$. We denote by $\mathbb{R}_{\exp}$ the structure of real numbers equipped with the usual arithmetic operations and the exponentiation function $x \mapsto e^x$. We will be interpreting formulas of $\mathcal{L}_e$ in $\mathbb{R}_{\exp}$.

The theory $\mathrm{Th}(\mathbb{R}_{\exp})$ does not admit quantifier elimination [93][1]. However, in

[1]van den Dries in 1982 showed that $y > 0 \wedge \exists w\ (wy = x \wedge z = ye^w)$ is not equivalent to a quantifier-free $\mathcal{L}_e$-formula.

1996, Macintyre and Wilkie showed that $\mathrm{Th}(\mathbb{R}_{\exp})$ is decidable assuming the truth of *Schanuel's conjecture.*

First stated in the 1960s, Schanuel's conjecture is a unifying conjecture in transcendental number theory that generalizes many of the classical results in the field.

Conjecture 2.7 (Schanuel's conjecture [58])**.** *If $\alpha_1, \ldots, \alpha_k \in \mathbb{C}$ are rationally linearly independent, then some k-element subset of $\{\alpha_1, \ldots, \alpha_k, e^{\alpha_1}, \ldots, e^{\alpha_k}\}$ is algebraically independent.*

Theorem 2.8 (Macintyre and Wilkie [63, 98])**.** *Assume Schanuel's conjecture for real numbers only. Then the theory $\mathrm{Th}(\mathbb{R}_{\exp})$ is decidable.*

Macintyre also claimed conditional decidability for a more general theory than $\mathrm{Th}(\mathbb{R}_{\exp})$ [62]. One may add functions $\sin_{\restriction[0,n]}$ and $\cos_{\restriction[0,n]}$ to the language $\mathcal{L}_e$ and consider the structure $\mathbb{R}_{\exp,\, \sin_{\restriction[0,n]},\, \cos_{\restriction[0,n]}}$ of real numbers with the usual arithmetic operations, the exponentiation function, and the sine and cosine functions restricted to the interval $[0, n]$. Note that it is necessary to restrict the sine and cosine functions to a bounded interval to get a decidable theory. The unrestricted functions would allow one to define the set of integers within the theory and perform quantification over this set, which would make the theory undecidable (by Godel).

The proof of this theorem was not published. A recent effort verified the proof, but unfortunately, that paper has not been published yet either.

Theorem 2.9 (Bounded Trig Conditional Decidability)**.** *Assume Schanuel's conjecture for complex numbers. For any $n \in \mathbb{N}$, the theory $\mathrm{Th}(\mathbb{R}_{\exp,\, \sin_{\restriction[0,n]},\, \cos_{\restriction[0,n]}})$ is decidable.*

We will need another result about $\mathbb{R}_{\exp}$ due to Wilkie. Note that this result is not conditional on Schanuel's conjecture.

Theorem 2.10 (O-minimality of $\mathbb{R}_{\exp}$ [97])**.** *The structure $\mathbb{R}_{\exp}$ is* o-minimal*, meaning that every set S definable in $\mathbb{R}_{\exp}$ has finitely many connected components in the Euclidean topology.*

2.4.3 Elementary numbers

An elementary point is an element of $\mathbb{C}^n$ that arises as an isolated, nonsingular solution of n equations in n variables $x_1, \ldots, x_n$, with each equation being either of the form $P(x_1, \ldots, x_n) = 0$, where $P \in \mathbb{Q}[x_1, \ldots, x_n]$ is a polynomial with rational coefficients, or of the form $x_j - e^{x_i} = 0$ for $i, j \in \{1, \ldots, n\}$. An elementary number is the polynomial image of an elementary point.

Roughly speaking, an elementary number is obtained by starting with the rationals and applying addition, subtraction, multiplication, division, exponentiation, and taking natural logarithms. Elementary numbers may be transcendental and, unsurprisingly, it is non-trivial to determine whether an elementary number is equal to zero.

Proposition 2.11 (Richardson [83]). *The problem of determining zeroness of an elementary number is semi-decidable. The problem is moreover decidable if one assumes Schanuel's conjecture.*

2.5 Semialgebraic Sets

A *semialgebraic* set is a subset of $\mathbb{R}^d$ definable (without parameters) in $\mathbb{R}_0$. By quantifier elimination, it is the solution set of a boolean combination of polynomial equalities and inequalities. We say that a function $\varphi : \mathbb{R}^l \to \mathbb{R}^m$ is semialgebraic if its graph is a semialgebraic subset of $\mathbb{R}^{l+m}$. Intuitively, semialgebraic functions are exactly the functions that can be specified using arithmetic and logical operations over the real numbers.

We say that a quantifier-free semialgebraic set S has *description size*[2] at most (n, d, τ) if it can be expressed by a boolean combination of polynomial equations and inequalities $P(x_1, \ldots, x_n) \bowtie 0$ with $\bowtie \in \{\leq, =\}$, involving polynomials $P \in \mathbb{Z}[x_1, \ldots, x_n]$ in at most n variables of total degree at most d with integer coefficients bounded in bitsize by τ.

Recall that by Tarski's theorem, we can apply quantifier elimination to semialgebraic sets, but this increases the size of the description. We will need the following singly exponential quantifier elimination result given in [5]. For a historical overview on this type of result see [5, Chapter 14, Bibliographical Notes].

Theorem 2.12 ([5, Theorem 14.16]). *Let $S \subseteq \mathbb{R}^{k+n_1+\cdots+n_\ell}$ be a quantifier-free semialgebraic set of description size at most $(k+n_1+n_2+\cdots+n_\ell, d, \tau)$. Let $Q_1, \ldots, Q_\ell \in \{\exists, \forall\}$ be a sequence of alternating quantifiers. Consider the set $S' \subseteq \mathbb{R}^k$ of all*

[2] The number of inequalities and equations is often relevant to the complexity of algorithmic operations on semialgebraic sets, but we will compute only size bounds on the set, so it is not relevant in this thesis.

$(x_1, \dots, x_k) \in \mathbb{R}^k$ *satisfying the first-order formula*

$$(Q_1(x_{1,1}, \dots, x_{1,n_1})) \dots (Q_\ell(x_{\ell,1}, \dots, x_{\ell,n_\ell})) \, . \\ ((x_1, \dots, x_k, x_{1,1}, \dots, x_{1,n_1}, \dots, x_{\ell,1}, \dots, x_{\ell,n_\ell}) \in S)$$

Then there exists a quantifier-free description of S' which has description size at most $(k, d^{O(n_1 \cdot \dots \cdot n_\ell)}, \tau d^{O(n_1 \cdot \dots \cdot n_\ell \cdot k)})$.

The next theorem is due to Vorobjov [96]. See also [43, Lemma 9] and [6, Theorem 4].

Theorem 2.13. *There exists an integer function* $\mathrm{Bound}(n, d, \tau) \in 2^{\tau d^{O(n)}}$ *with the following property:*

Let K be a compact semialgebraic set of description size at most (n, d, τ). Then K is contained in a ball centred at the origin of radius at most $\mathrm{Bound}(n, d, \tau)$.

A closely related result, due to [52], yields a lower bound on the minimum of a polynomial over a compact semialgebraic set, provided the minimum is non-zero. The result in [52] mentions explicit constants, which is more than we need.

Theorem 2.14 ([52, Theorem 1])**.** *There exists an integer function* $\mathrm{LowerBound}(n, d, \tau) \in 2^{(\tau d)^{n^{O(1)}}}$ *such that the following holds true:*

Let $P \in \mathbb{Q}[x_1, \dots, x_n]$ be a polynomial of degree at most d, whose coefficients have bitsize at most τ. Let K be a compact semialgebraic set of description size at most (n, d, τ). If $\min_{x \in K} P(x) > 0$ then $\min_{x \in K} P(x) > 1/\mathrm{LowerBound}$.

With the help of Theorem 2.12, Theorem 2.14 can be generalised to yield a lower bound on the distance of two disjoint compact semialgebraic sets. A very similar result is proved in [86] under more general assumptions. Unfortunately, the bound stated there is not sufficiently fine-grained for our purpose, since the author do not distinguish the dimension of a set from the other description size parameters.

Lemma 2.15. *There exists an integer function* $\mathrm{Sep}(n, d, \tau) \in 2^{(\tau d)^{n^{O(1)}}}$ *with the following property:*

Let K and L be compact semialgebraic sets of description size at most (n, d, τ). Assume that every $x \in K$ has positive euclidean distance to L. Then $\inf_{x \in K} d(x, L) > 1/\mathrm{Sep}(n, d, \tau)$.

Proof. If either K or L are empty then the result is trivial. Thus, let us assume that both sets are non-empty.

Consider the semialgebraic set

$$S = \left\{(x, y) \in K \times L \mid \forall z \in L. \left(\|x - z\|_2^2 \geq \|x - y\|_2^2\right)\right\}.$$

By Theorem 2.12 the set S has description size $(2n, d^{n^{O(1)}}, \tau d^{n^{O(1)}})$. By compactness, the distance $\mathrm{dist}_{\ell^2}(x, L)$ is attained in a point $y \in L$ for all $x \in K$, so that for all $x \in K$ there exists $y \in L$ such that $(x, y) \in S$.

We clearly have

$$\inf_{x \in K} \mathrm{dist}_{\ell^2}(x, L)^2 = \inf_{(x,y) \in S} \|x - y\|_2^2. \tag{2.3}$$

The right-hand side of (2.3) is a polynomial, so the result follows from Theorem 2.14. □

2.6 Diophantine approximation

The analysis of problems about linear dynamical systems often reduces to that of the orbit $\langle \Gamma^n \mid n \in \mathbb{N} \rangle$ where $\Gamma^n = (\gamma_1^n, \ldots, \gamma_k^n)$ for $\gamma_1, \ldots, \gamma_k \in \mathbb{T}$. Let $\mathcal{T} = \mathrm{cl}(\{\Gamma^n : n \in \mathbb{N}\})$ be the topological closure of this discrete orbit. The set $\mathcal{T}$ is semialgebraic and well-understood with the help of Kronecker's theorem in simultaneous Diophantine approximation [48].

Theorem 2.16 (Kronecker)**.** *Let $\theta_1, \ldots, \theta_k, \varphi_1, \ldots, \varphi_k \in \mathbb{R}$ be such that for any $a_1, \ldots, a_k \in \mathbb{Z}$,*

$$\sum_{i=1}^{k} a_i \theta_i \in \mathbb{Z} \Rightarrow \sum_{i=1}^{k} a_i \varphi_i \in \mathbb{Z}.$$

For any $\epsilon > 0$ there exist infinitely many $n \in \mathbb{N}$ such that $\{n\theta_i - \varphi_i\} < \epsilon$ for all $1 \leq i \leq k$, where $\{x\}$ denotes the distance from $x \in \mathbb{R}$ to the nearest integer.

To apply this theorem to our situation, let

$$\mathcal{T} = \{(z_1, \ldots, z_k) : \forall a_1, \ldots, a_k \in \mathbb{Z} : \gamma_1^{a_1} \cdots \gamma_k^{a_k} = 1 \Rightarrow z_1^{a_1} \cdots z_k^{a_k} = 1\}.$$

For $z = (z_1, \ldots, z_k) \in \mathcal{T}$, by considering $\theta_i = \frac{\arg(\gamma_i)}{2\pi}$ and $\varphi_i = \frac{\arg(z_i)}{2\pi}$ for $1 \leq i \leq k$ we can deduce that for each $\epsilon > 0$ there exists n such that $||z - \Gamma^n|| < \epsilon$ and hence the orbit $\langle \Gamma^n \mid n \in \mathbb{N} \rangle$ is dense in $\mathcal{T}$. On the other hand, using Masser's deep results [66] (Theorem 2.5) about multiplicative relations between algebraic numbers one can compute, in polynomial time, a finite basis for $\{(a_1, \ldots, a_k) \in \mathbb{Z}^k : \gamma_1^{a_1} \cdots \gamma_k^{a_k} = 1\}$.

Hence $\mathcal{T}$ is closed, semialgebraic and effectively computable. It then follows that $\mathcal{T} = \mathrm{cl}(\{\Gamma^n : n \in \mathbb{N}\})$.

We will also need the following lemma which is a consequence of the effective computability of $\mathcal{T} = \mathrm{cl}(\{\Gamma^n : n \in \mathbb{N}\})$ as a semialgebraic set.

Corollary 2.17. *Let $R = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_k) \in \mathbb{R}^{2k\times 2k}$ be a block diagonal matrix where Λ_i is an algebraic rotation matrix for $1 \leq i \leq k$. The closure of the set $\{R^n x : n \in \mathbb{N}\}$, for x with algebraic entries, is semialgebraic and effectively computable.*

The proof follows immediately from diagonalising R^n and observing that all eigenvalues of R are algebraic numbers in $\mathbb{T}$.

We will also need a quantitative version of the theorem due to Eike Neumann ([32, Corollary 13, Appendix D]) that tells us how long we need to wait for the orbit to come close to a given point.

Theorem 2.18 (Quantitative Kronecker [32]). *There exists a function $\mathrm{Kron}(n, \tau, P) \in 2^{(\tau P)^{n^{O(1)}}}$, such that the following holds true:*

Let $\lambda_1, \ldots, \lambda_n$ be algebraic numbers of modulus 1 and description size $(1, n, \tau)$. Let $\alpha_1, \ldots, \alpha_n$ be complex numbers in the closure of the sequence $(\lambda_1^t, \ldots, \lambda_n^t)_{t\in\mathbb{N}}$. Then for all $P \in \mathbb{N}$ there exists a $t \leq \mathrm{Kron}(n, \tau, P)$ such that $|\alpha_j - \lambda_j^t| < 2^{-P}$ for all $j \in \{1, \ldots, n\}$.

2.7 Description size bounds on converting a semialgebraic set to a new basis

When working with iterated matrix maps it is often useful to transform the system under consideration so that the matrix is in Jordan form. We now combine results from the previous sections to control the description size increase that may occur when transforming a semialgebraic set to a new basis.

The following example illustrates the kind of blowup that may occur: Consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 + 10^{-2} \end{bmatrix},$$

its associated Jordan decomposition

$$A = QJQ^{-1} = \begin{bmatrix} 1 & 0 & 10000 \\ 0 & 1 & 100 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1.01 \end{bmatrix} \begin{bmatrix} 1 & 0 & -10000 \\ 0 & 1 & -100 \\ 0 & 0 & 1 \end{bmatrix}$$

and the (singleton) set $S = \{(0,0,1) \in \mathbb{R}^3\}$. In the Jordan basis, this set becomes $\{(-10^4, -10^2, 1) \in \mathbb{R}^3\}$, increasing the bitsize of the coefficients defining the set by a factor polynomial in the dimension and the bitsize of the matrix entries.

Let A and A' be matrices (not necessarily of the same dimensions) and K and K' be semialgebraic sets. We say (A, K) and (A', K') are *equivalent systems* if there exists a bijective map B such that $A^k x \in K \iff A'^k B(x) \in K'$ for all $k \in \mathbb{N}$. Note that equivalent systems have the same escape properties.

Normally if A' is the Jordan normal form of A via a change of basis Q (i.e. $A = QA'Q^{-1}$) we would expect Q^{-1} to serve as the map B. However a naive approach to defining the set $Q^{-1}K$ via quantifier elimination may result in significantly greater description size than K itself, so in fact we make a more general definition. This lets us introduce additional dimensions to safely manipulate the algebraic numbers that arise in the Jordan normal form.

Lemma 2.19. *Let $A \in \mathbb{Q}^{n\times n}$ be a matrix with entries of bitsize a and K be a semi-algebraic set of description size (n, d, τ).*

Then there exists $m \le n$ and a system (J', K') equivalent to (A, K) such that the matrix $J' \in (\overline{\mathbb{Q}} \cap \mathbb{R})^{(n+m)\times(n+m)}$ is in real Jordan normal form and K' has description size $(n + m, n^2 d, \tau + d \log 2n^3 + d\sigma)$, where $\sigma = O^*(an^3)$.

Proof. By subsection 2.3.1 we can compute in polynomial time real algebraic numbers $\gamma_1, \dots, \gamma_m$ and a matrix $Q \in \mathbb{Q}(\gamma_1, \dots, \gamma_m)^{n\times n}$ such that $A = QJQ^{-1}$, where J is in real Jordan normal form. Note that $m \le n$ since all complex eigenvalues occur in conjugate pairs. Let δ be a bound on the degrees of $\gamma_1, \dots, \gamma_m$. Since the real and imaginary parts may be obtained by adding or subtracting conjugates, we have $\delta \le n^2$ by elementary Galois theory.

More precisely, we can compute in polynomial time:

1. Univariate polynomials $f_1, \dots, f_m$ with integer coefficients such that $f_j(\gamma_j) = 0$ for all $j = 1, \dots, m$.

2. Rational numbers $a_1, b_1, \dots, a_m, b_m$, such that γ_j is the unique root of f_j in the interval $[a_j, b_j]$..

3. For $i = 1, \dots, n$ and $j = 1, \dots, n$ polynomials of degree at most δ $Q_{i,j} \in \mathbb{Q}[x]$ and indexes $\ell_{i,j}$ such that the matrix Q at row i and column j is given by the algebraic number $Q_{i,j}(\gamma_{\ell_{i,j}})$.

Let $\sigma \in \mathbb{N}$ be a common bound on the following quantities:

1. The bitsize of the coefficients of $f_1, \ldots, f_m$.

2. The bitsize of the endpoints of the isolating intervals $[a_j, b_j]$.

3. The bitsize of the coefficients of the polynomials $Q_{j,k}$.

Then σ is computable in polynomial time from A, so it depends polynomially on n and on a (the bitsize of the entries of A). From the proof of Theorem 2.6 we can derive that in fact $\sigma = O^*(an^3)$ (suppressing logarithmic factors in n).

We fix $K' = (Q^{-1}K) \times \{\gamma_1, \ldots, \gamma_m\}$ and note that $A^k x \in K$ if and only if

$$(J \times I_m)^k (Q^{-1}x, (\gamma_1, \ldots, \gamma_m)) \in (Q^{-1}K) \times \{(\gamma_1, \ldots, \gamma_m)\}.$$

Here the map $x \mapsto (Q^{-1}x, (\gamma_1, \ldots, \gamma_m))$ is the map B required for equivalence. The last m coordinates are added to allow us to manipulate these algebraic numbers within the description of K' symbolically through a formula. (If we did not do this, and tried to apply quantifier elimination to K' to work directly with the set $Q^{-1}K$, we would end up with a set that has description size $(n, (dn)^{O(n)}, (\tau + d\sigma)(dn)^{O(n^2)})$ which is exponential in n.)

We now show there exists a description of the set K' with the claimed size.

Let $\Phi(x_1, \ldots, x_n)$ be the formula that describes K. We introduce fresh variables $z_1, \ldots, z_m$ and consider the formula

$$\Psi(z_1, \ldots, z_m) \wedge \widehat{\Phi}(x_1, \ldots, x_n, z_1, \ldots, z_m)$$

where $\Psi(z_1, \ldots, z_m)$ is the conjunction of the terms

$$f_j(z_j) = 0 \wedge z_j \geq a_j \wedge z_j \leq b_j$$

which ensures $z_j = \gamma_j$ for $j = 1, \ldots, m$, and $\widehat{\Phi}$ is obtained from Φ by replacing each atom $P(x_1, \ldots, x_n) \bowtie 0$ in Φ by the atom

$$P\left(\sum_{k=1}^{n} Q_{1,k}(z_{\ell_{1,k}})x_k, \ldots, \sum_{k=1}^{n} Q_{n,k}(z_{\ell_{n,k}})x_k,\right) \bowtie 0.$$

It is not hard to see that this new formula describes the set K'. Evidently, the number of variables in this description is $n + m$. The formula Ψ involves $3m$ polynomials of degree at most δ whose coefficients are bounded in bitsize by σ.

It remains to determine the description size of the formula $\widehat{\Phi}$. We claim that the degrees of the polynomials in $\widehat{\Phi}$ are bounded by $\delta \cdot d$ and that the bitsize of their coefficients is bounded by $\tau + d(\log(n) + \log(\delta + 1) + \sigma)$. This is established by a straightforward but cumbersome calculation. We recall the multinomial theorem:

Lemma 2.20 (Multinomial theorem)**.** *Let R be a ring. Let N be a positive integer. Let $z_1, \ldots, z_N \in R$. Then*

$$\left(\sum_{k=1}^{N} z_k\right)^e = \sum_{j_1+\cdots+j_N=e} \binom{e}{j_1, \ldots, j_N} \prod_{t=1}^{N} z_t^{j_t},$$

where

$$\binom{e}{j_1, \ldots, j_N} = \frac{e!}{j_1! \cdot \cdots \cdot j_N!}.$$

It will be convenient to make use of the following straightforward application of the distributivity law of multiplication over addition. Here $\coprod$ denotes the disjoint union.

Lemma 2.21. *Let I be a finite set. Let J be a set-valued function that sends each $i \in I$ to a finite set $J(i)$. Let R be a ring. For all $(i,j) \in I \times \coprod_{i\in I} J(i)$, let $a_{i,j}$ be an element of R. Then we have*

$$\prod_{i\in I} \sum_{j\in J(i)} a_{i,j} = \sum_{f} \prod_{i\in I} a_{i,f(i)}$$

where f ranges over all functions

$$f\colon I \to \coprod_{i\in I} J(i)$$

satisfying $f(i) \in J(i)$ for all $i \in I$.

Now, let $P(x_1, \ldots, x_n) \bowtie 0$ be an atom in Φ. This atom is a sum of monomials

$$C \cdot x_1^{e_1} \cdot \cdots \cdot x_n^{e_n}$$

with $\log |C| \leq \tau$ and $e_1 + \cdots + e_n \leq d$. It suffices to bound the degrees and the bitsize of the coefficients of the polynomials that are obtained by applying our substitution of variables to monomials of this form.

Under our substitution such a monomial becomes:

$$C \cdot \left(\sum_{j=1}^{n} Q_{1,j}(z_{\ell_{1,j}})x_j\right)^{e_1} \cdot \cdots \cdot \left(\sum_{j=1}^{n} Q_{n,j}(z_{\ell_{n,j}})x_j\right)^{e_n} = C \prod_{k=1}^{n} \left(\sum_{j=1}^{n} Q_{k,j}(z_{\ell_{k,j}})x_j\right)^{e_k}.$$

Apply the multinomial theorem to the expressions $\left(\sum_{j=1}^{n} Q_{k,j}(z_{\ell_{k,j}})x_j\right)^{e_k}$ to obtain:

$$C \cdot \prod_{k=1}^{n} \left(\sum_{j_{k,1}+\cdots+j_{k,n}=e_k} \binom{e_k}{j_{k,1}, \ldots, j_{k,n}} \prod_{t=1}^{n} \left(Q_{k,t}(z_{\ell_{k,t}})x_t\right)^{j_{k,t}} \right).$$

Write

$$Q_{k,t}(z_{\ell_{k,t}}) = \sum_{p=0}^{\delta} \alpha_{k,t,p} z_{\ell_{k,t}}^{p}.$$

Applying the multinomial theorem to the terms

$$\left(Q_{k,t}(z_{\ell_{k,t}})x_t\right)^{j_{k,t}} = \left(\sum_{p=0}^{\delta} \alpha_{k,t,p} z_{\ell_{k,t}}^{p} x_t\right)^{j_{k,t}}$$

we obtain

$$\left(Q_{k,t}(z_{\ell_{k,t}})x_t\right)^{j_{k,t}} = \sum_{r_0+\cdots+r_\delta=j_{k,t}} \binom{j_{k,t}}{r_0,\ldots,r_\delta} \prod_{s=0}^{\delta} \alpha_{k,t,s}^{r_s} z_{\ell_{k,t,s}}^{sr_s} x_t^{r_s}.$$

The full expression is hence:

$$C \cdot \prod_{k=1}^{n} \left(\sum_{j_{k,1}+\cdots+j_{k,n}=e_k} \binom{e_k}{j_{k,1},\ldots,j_{k,n}} \prod_{t=1}^{n} \sum_{r_0+\cdots+r_\delta=j_{k,t}} \binom{j_{k,t}}{r_0,\ldots,r_\delta} \prod_{s=0}^{\delta} \alpha_{k,t,s}^{r_s} z_{\ell_{k,t,s}}^{sr_s} x_t^{r_s} \right).$$

Write this as:

$$C \cdot \prod_{k=1}^{n} \sum_{j_{k,1}+\cdots+j_{k,n}=e_k} \binom{e_k}{j_{k,1},\ldots,j_{k,n}} \prod_{t=1}^{n} \sum_{r_0+\cdots+r_\delta=j_{k,t}} c_{k,j_{k,1},\ldots,j_{k,n},t,r_0,\ldots,r_\delta}.$$

Apply Lemma 2.21 to move out the innermost sum, thus obtaining an equal expression:

$$C \cdot \prod_{k=1}^{n} \left(\sum_{j_{k,1}+\cdots+j_{k,n}=e_k} \sum_{f} \binom{e_k}{j_{k,1},\ldots,j_{k,n}} \prod_{t=1}^{n} c_{k,j_{k,1},\ldots,j_{k,n},t,f(t)} \right).$$

where the sum $\sum_f$ ranges over all functions $f\colon \{1,\ldots,n\} \to \mathbb{N}^{\delta+1}$ with $f(t) = (r_0,\ldots,r_\delta)$ satisfying $r_0+\cdots+r_\delta = j_{k,t}$.

Write the result as:

$$C \cdot \prod_{k=1}^{n} \sum_{j_{k,1}+\cdots+j_{k,n}=e_k} \sum_{f} d_{k,j_{k,1},\ldots,j_{k,n},f}.$$

Apply Lemma 2.21 again to obtain that this is equal to:

$$\sum_{g} C \cdot \prod_{k=1}^{n} d_{k,g(k)},$$

where g ranges over all functions $g\colon \{1,\ldots,n\} \to \mathbb{N}^n \times (\mathbb{N}^{\delta+1})^{\{1,\ldots,n\}}$ with $g(k) = (j_{k,1},\ldots,j_{k,n},f)$ satisfying $j_{k,1}+\cdots+j_{k,n} = e_k$ and f as above.

Thus, the final result is a sum of monomials of the form

$$
\begin{aligned}
C \cdot \prod_{k=1}^{n} d_{k,g(k)} &= C \cdot \prod_{k=1}^{n} \binom{e_k}{j_{k,1}, \dots, j_{k,n}} \prod_{t=1}^{n} c_{k,j_{k,1},\dots,j_{k,n},t,f(t)} \\
&= C \cdot \prod_{k=1}^{n} \binom{e_k}{j_{k,1}, \dots, j_{k,n}} \prod_{t=1}^{n} \binom{j_{k,t}}{r_0(t), \dots, r_\delta(t)} \prod_{s=0}^{\delta} \alpha_{k,t,s}^{r_s(t)} z_{\ell_{k,t,s}}^{sr_s(t)} x_t^{r_s(t)},
\end{aligned}
$$

Where $j_{k,1} + \cdots + j_{k,n} = e_k$ and $r_0(t), \dots, r_\delta(t)$ are functions of t satisfying $r_0(t) + \cdots + r_\delta(t) = j_{k,t}$.

Since $e_1 + \cdots + e_n \leq d$, the degrees of these monomials are bounded by $\delta \cdot d$.

Let us compute a bound on the bitsize of the coefficients. We have:

$$
\begin{aligned}
&\log\left(|C| \cdot \prod_{k=1}^{n} \binom{e_k}{j_{k,1}, \dots, j_{k,n}} \prod_{t=1}^{n} \binom{j_{k,t}}{r_0(t), \dots, r_\delta(t)} \prod_{s=0}^{\delta} |\alpha_{k,t,s}|^{r_s(t)} \right) \\
&\leq \tau + \sum_{k=1}^{n} \log \binom{e_k}{j_{k,1}, \dots, j_{k,n}} + \sum_{k=1}^{n}\sum_{t=1}^{n} \log \binom{j_{k,t}}{r_0(t), \dots, r_\delta(t)} + \sum_{k=1}^{n}\sum_{t=1}^{n}\sum_{s=0}^{\delta} r_s(t)\sigma.
\end{aligned}
$$

Use the estimate $\binom{f}{k_1,\dots,k_m} \leq m^f$ to obtain:

$$
\begin{aligned}
&\log\left(|C| \cdot \prod_{k=1}^{n} \binom{e_k}{j_{k,1}, \dots, j_{k,n}} \prod_{t=1}^{n} \binom{j_{k,t}}{r_0(t), \dots, r_\delta(t)} \prod_{s=0}^{\delta} |\alpha_{k,t,s}|^{r_s(t)} \right) \\
&\leq \tau + \sum_{k=1}^{n} e_k \log(n) + \sum_{k=1}^{n}\sum_{t=1}^{n} j_{k,t} \log(\delta+1) + \sum_{k=1}^{n}\sum_{t=1}^{n}\sum_{s=0}^{\delta} r_s(t)\sigma. \\
&\leq \tau + d\log(n) + d\log(\delta+1) + d\sigma. \\
&= \tau + d(\log(n) + \log(\delta+1) + \sigma).
\end{aligned}
$$

Thus, everything is shown.

□

2.8 Linear Recurrence Sequences

A *semiring* is a ring without additive inverses. A sequence $\boldsymbol{u} = (u_n)_{n=0}^{\infty}$ of elements of a semiring K is called *K-rational* if there exists $d \geq 1$, $c, s \in K^d$ and $M \in K^{d\times d}$ such that $u_n = c^\top M^n s$ for all n. When K is a field, a sequence is K-rational if and only if it satisfies a linear recurrence relation

$$u_n = a_1 u_{n-1} + \cdots a_d u_{n-d} \quad (n \geq d)$$

where $a_1, \dots, a_d \in K$. In this case we also call $\boldsymbol{u}$ a *linear recurrence sequence (LRS).*

It is easy to see that an LRS is K-rational via its companion matrix

$$C := \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_d & a_{d-1} & a_{d-2} & \cdots & a_1 \end{bmatrix} \in K^{d\times d}.$$

The converse follows from the Cayley-Hamilton theorem.

With the unique minimal order recurrence satisfied by $\boldsymbol{u}$ we associate the *characteristic polynomial*

$$P(X) = X^d - a_1 X^{d-1} - \cdots - a_d .$$

The roots of $P(X)$ are called the *characteristic roots* of $\boldsymbol{u}$. Writing $\lambda_1, \ldots, \lambda_m$ for the distinct characteristic roots, in non-increasing order of absolute value, the sequence $\boldsymbol{u}$ admits a closed-form *exponential-polynomial* representation:

$$u_n = \sum_{i=1}^{m} P_i(n)\lambda_i^n ,$$

where the P_i are univariate polynomials whose coefficients are algebraic over K [45].

We say that $\boldsymbol{u}$ is *non-degenerate* if no quotient of two distinct characteristic roots is a root of unity. We also say that a matrix $M \in \mathbb{Q}^{d\times d}$ is non-degenerate if no quotient of two distinct eigenvalues is a root of unity.

We can always split a linear recurrence sequence into exponentially many non-degenerate subsequences.

Theorem 2.22 ([39, Theorem 1.2])**.** *For every LRS $\boldsymbol{u}$ of order k there exists an integer $L = 2^{O(k\sqrt{\log k})}$ such that the sequences $(u_{nL+r})_{n\in\mathbb{N}}$ for $0 \le r < L$ are all either identically zero or non-degenerate.*

In this thesis we exclusively consider sequences with rational entries. We say that an LRS $\boldsymbol{u}$ is *dominated* if λ_1 is the unique characteristic root of maximum modulus. Note that in this case λ_1 is necessarily real. We have the following straightforward propositions concerning dominated LRS.

Proposition 2.23. *If an LRS $\boldsymbol{u}$ is dominated then $\{n \in \mathbb{N} : u_n \ge 0\}$ is an effectively computable ultimately periodic set.*

Proof. Consider the closed form representation $u_n = \sum_{i=1}^{m} P_i(n)\lambda_i^n$ with unique dominant root λ_1, necessarily real. Suppose that P_1 has degree k and leading coefficient a. Then we have $\frac{u_n}{n^k|\lambda_1|^n} = a(\lambda_1/\,|\lambda_1|)^n + o(1)$. Hence for sufficiently large n the sign of u_n is determined by the sign of a and the parity of n. □

Proposition 2.24. *If $\boldsymbol{u}$ is a non-degenerate LRS such that some subsequence $(u_{cn+d})_{n=0}^{\infty}$ is dominated, where c is a positive integer and $d \in \{0, 1 \ldots, c-1\}$, then $\boldsymbol{u}$ itself is dominated.*

Proof. The sequence $\boldsymbol{u}$ admits a closed-form representation $u_n = \sum_{i=1}^{m} P_i(n)\lambda_i^n$, where $\lambda_1, \ldots, \lambda_m$ are the characteristic roots and $P_1, \ldots, P_m$ are polynomials. Then

$$\begin{aligned} u_{cn+d} &= \sum_{i=1}^{m} P_i(cn+d)\lambda_i^{cn+d} \\ &= \sum_{i=1}^{m} Q_i(n)(\lambda_i^c)^n \end{aligned}$$

where $Q_i(n) := P_i(cn+d)\lambda_i^d$ for $i \in \{1, \ldots, m\}$.

Note that the polynomials $Q_1, \ldots, Q_m$ are non-zero and, by non-degeneracy of $\boldsymbol{u}$, the numbers $\lambda_1^c, \ldots, \lambda_m^c$ are pairwise distinct. Since the sequence $(u_{cn+d})_{n=0}^{\infty}$ is dominated, we have that λ_1^c is its unique characteristic root of maximum modulus. But then λ_1 is the unique characteristic root of $\boldsymbol{u}$ of maximum modulus. □

2.8.1 Problems on Linear Recurrence Sequences

We say that a hyperplane or a halfspace is rational if it can be defined in the form $\{x \in \mathbb{R}^d \mid c^\top x \bowtie 0\}$ where $\bowtie \in \{\geq, >, =\}$ and $c \in \mathbb{Q}^d$.

2.8.1.1 Skolem

The *Skolem problem* over $\mathbb{Q}$ is to decide, given a rational LRS $(u_n)_{n\in\mathbb{N}}$, whether $u_n = 0$ for some n.

Writing $u_n = c^\top M^n s$, $u_n = 0$ for some n if and only if the orbit of (M, s) reaches the rational hyperplane $H = \{x \mid c^\top x = 0\}$. Conversely, the orbit of (M, s) reaches a rational hyperplane H with normal vector $c \in \mathbb{Q}^d$ if and only if the LRS $u_n = c^\top M^n s$ has a zero. Therefore, the Skolem Problem is equivalent to the Reachability Problem with rational hyperplane targets: given $(M, s) \in \mathbb{Q}^{d\times d} \times \mathbb{Q}^d$ and a rational hyperplane $H \subseteq \mathbb{R}^d$, decide whether the orbit of (M, s) reaches H.

It is decidable in dimension upto 4, with complexity $\mathbf{NP}^{\mathbf{RP}}$ [20]. No unconditional results are known beyond order 4.

Conditionally, Bilu et al [14] showed decidability for the Skolem problem with *diagonalisable* matrices, assuming the p-adic Schanuel conjecture and the Skolem Conjecture (aka Exponential Local-Global Principle) in order to guarantee termination of their algorithm. This (conditionally) also solves the last open case of the Skolem problem for dimension 5.

A new direction of work considers *Universal Skolem Sets*: subsets S of the positive integers such that it is decidable whether a given non-degenerate linear recurrence sequence has a zero in S. The Skolem problem is equivalent to showing $\mathbb{N}$ itself is a Universal Skolem Set. There is a known Universal Skolem Set of lower density at least 0.29 [61].

2.8.1.2 Positivity

The *Positivity problem* over $\mathbb{Q}$ is to decide, given a rational LRS $(u_n)_{n\in\mathbb{N}}$, whether $u_n \geq 0$ for all n.

By the same argument as before, it is equivalent to deciding, given a linear dynamical system (M, s) and a rational halfspace $H = \{x \mid c^\top x \geq 0\}$, whether the orbit of (M, s) always remains in H.

The Positivity Problem is decidable for LRS in dimension 5 or less, with complexity in $\mathbf{coNP}^{\mathbf{PP}^{\mathbf{PP}^{\mathbf{PP}}}}$, i.e., within the fourth level of the Counting Hierarchy [74].

The decidability of either Positivity or Ultimate Positivity for LRS in dimension 6 would entail major breakthroughs in analytic number theory (certain open problems in Diophantine approximation of transcendental numbers long believed to be hard would become solvable) [74].

2.8.1.3 Ultimate Positivity

The *Ultimate Positivity problem* over $\mathbb{Q}$ is to decide, given a rational LRS $(u_n)_{n\in\mathbb{N}}$, whether $u_n \geq 0$ for all sufficiently large n.

It is equivalent to deciding, given a linear dynamical system (M, s) and a rational halfspace $H = \{x \mid c^\top x \geq 0\}$, whether the orbit of (M, s) is contained in H apart from finitely many points.

The Ultimate Positivity Problem is decidable for LRS in dimension 5 or less in polynomial time [74]. For diagonalisable matrices, Ultimate Positivity is decidable in all dimensions, with complexity between **coNP** and **PSPACE** [75].

2.8.1.4 Scalar Reachability (Hyperplane Reachability)

Scalar Reachability (also known as the *Hyperplane Reachability problem*) generalizes the Skolem problem to multiple matrices. The *Scalar Reachability problem* over $\mathbb{N}$ is to decide, given a set of matrices $A_1, \ldots, A_k \in \mathbb{N}^{d\times d}$, vectors $u, v \in \mathbb{N}^d$ and a scalar $\lambda \in \mathbb{N}$, whether there exists A in the semigroup generated by $A_1, \ldots, A_k$ such that $u^\top A v = \lambda$.

There is an undecidable instance of Scalar Reachability using 5 integer 3×3 matrices and the scalar 0 [46].

2.8.1.5 Halfspace Reachability

The *Halfspace Reachability problem* over $\mathbb{Q}$ is to decide, given a set of matrices $A_1, \ldots, A_k \in \mathbb{Q}^{d \times d}$, vectors $u, v \in \mathbb{Q}^d$ and a scalar $\lambda \in \mathbb{Q}$, whether there exists A in the semigroup generated by $A_1, \ldots, A_k$ such that $u^\top A v \geq \lambda$.

Halfspace Reachability is also undecidable. We provide an explicit proof, using a reduction from the Scalar Reachability problem over integers. The reduction generalizes the standard reduction from the Skolem problem over $\mathbb{N}$ of dimension d to the Positivity problem over $\mathbb{N}$ of dimension $d^2 + 1$.

We make essential use of the *mixed-product* property of the Kronecker product, namely that $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for matrices A, B, C, D of compatible dimensions.

Theorem 2.25 (Undecidability of Halfspace Reachability)**.** *The Halfspace Reachability problem over $\mathbb{Q}$ is undecidable for 5 rational 10×10 matrices and scalar $1/2$.*

Proof. We reduce from the Scalar Reachability problem over $\mathbb{N}$ with 5 integer 3×3 matrices and scalar 0.

Let $A_1, \ldots, A_5 \in \mathbb{N}^{3 \times 3}$, vectors $u, v \in \mathbb{N}^3$ and $\lambda = 0$ be the undecidable instance of Scalar Reachability constructed in [46]. We construct 10×10 rational matrices $B_1, \ldots, B_5$, vectors $x, y \in \mathbb{Q}^{10}$ and scalar $\mu = 1/2$ such that there exists a rational matrix B in the semigroup generated by $B_1, \ldots, B_5$ such that $x^\top B y \geq \mu$ if and only if there exists an integer matrix A in the semigroup generated by $A_1, \ldots, A_5$ such that $u^\top A v = \lambda$.

Given non-negative integers m, n, write $0^{m \times n}$ for the zero matrix of dimension $m \times n$. Define

$$\begin{aligned} x^\top &:= \begin{bmatrix} 1 & -u^\top \otimes u^\top \end{bmatrix}, \\ B_i &:= \begin{bmatrix} 1 & 0^{1 \times 9} \\ 0^{9 \times 1} & A_i \otimes A_i \end{bmatrix} \text{ for } i \in \{1, \ldots, 5\} \\ \text{and } y &:= \begin{bmatrix} 1 \\ v \otimes v \end{bmatrix}. \end{aligned}$$

Now let A be a product of the generators A_i and let B be constructed via generators B_i of corresponding indices. Observe (using the mixed-product property) that

$$
\begin{aligned}
& x^\top B y \geq 1/2 \\
\iff & \begin{bmatrix} 1 & -u^\top \otimes u^\top \end{bmatrix} \begin{bmatrix} 1 & 0^{1\times 9} \\ 0^{9\times 1} & A \otimes A \end{bmatrix} \begin{bmatrix} 1 \\ v \otimes v \end{bmatrix} \geq 1/2 \\
\iff & 1 + (-u^\top A v) \otimes (u^\top A v) \geq 1/2 \\
\iff & 1 - (u^\top A v)^2 \geq 1/2.
\end{aligned}
$$

Notice that $u^\top A v$ is an integer, and thus the above inequality holds if and only if $u^\top A v = 0$. This completes the reduction. $\square$

Chapter 3

Nonnegativity Problems for Matrix Semigroups

3.1 Introduction

In this chapter, we introduce the *Non-negative Membership Problem*, which asks to determine whether the semigroup generated by a given finite set of rational matrices contains a non-negative matrix, i.e., a matrix all of whose entries are non-negative. We show that this problem is undecidable in general but is decidable in the commutative case subject to Schanuel's conjecture. Our reliance on Schanuel's conjecture arises because we reduce the commutative case of the Non-negative Membership Problem to the decision problem for the first-order theory of real-closed fields with exponential. As shown by Macintyre and Wilkie [63], the latter theory is decidable if one assumes Schanuel's conjecture for the real exponential.

We call a matrix M *eventually non-negative* if for all but finitely many $n \in \mathbb{N}$ the matrix power M^n is non-negative. A key lemma in our main decidability result involves determining whether a given matrix is eventually non-negative. We give an effective characterisation of the set of non-negative powers of a matrix, answering a question posed by S. Akshay [2]. The characterisation is relatively straightforward and relies on classical results about rational sequences over the semi-ring of non-negative rational numbers. We also provide a standalone proof of this result via the Perron-Frobenius theorem.

Note that the problem of determining whether, for some fixed index (i, j), the sequence of scalars $\langle (M^n)_{i,j} : n \in \mathbb{N} \rangle$ is ultimately non-negative is equivalent to the Ultimate Positivity Problem for linear recurrence sequences, decidability of which is a longstanding open problem [74].

It is immediate that a matrix semigroup contains a non-negative matrix if and only if it contains an eventually non-negative matrix. Using a symbolic version of our criterion characterizing eventually non-negative matrices, we reduce the Non-negative Membership Problem to a version of integer programming with certain transcendental constants (namely logarithms of algebraic numbers). In turn we reduce the solution of such integer programs to the decision problem for the first-order theory of real closed fields with exponential.

As far as we are aware, the Non-negative Membership Problem has not been directly addressed before. We note however that decidability of the version of this problem for sub-semigroups of the group $\mathbf{GL}(2,\mathbb{Z})$ of 2×2 invertible integer matrices follows directly from [24, Theorem 13].

3.1.1 The Positive Membership Problem and related problems

A simpler variant of our main result concerns the problem of deciding whether a finitely generated matrix semigroup contains a positive matrix, i.e., a matrix all of whose entries are strictly positive. We show decidability for commuting matrices, leaving the general case open.

To show decidability in the commutative case, we rely on a known characterisation of eventually positive matrices, due to Noutsos [72]. While we still need to invoke Schanuel's conjecture in this case, we do so through the use of a procedure of Richardson [83] for deciding equality of elementary numbers (which is much more straightforward than the result of Macintyre and Wilkie mentioned above).

A related problem which has been intensively studied is the *Primitivity Problem*. A multiplicative semigroup of **nonnegative** matrices is called *primitive* if it possesses at least one (strictly) positive matrix. The primitivity problem asks whether a given set of (possibly noncommuting) matrices generates a primitive semigroup.

The restriction that the given matrices be non-negative causes significant differences from the problems we consider in this chapter. For one, since no cancellation of sums of negative and positive values is possible, only whether an entry of the matrix is zero or nonzero is relevant. Thus the primitivity problems for rational and integer matrices can be reduced to the problem over $(0,1)$-matrices.

By a result of Wielandt [87], if a single $n \times n$ nonnegative matrix A is primitive, then the specific power A^{n^2-2n+2} is positive (and so are all further powers). Thus primitivity for the semigroup generated by a single matrix is in PTIME. Surprisingly, the primitivity problem for two $n \times n$ matrices is PSPACE-complete [42]!

The upper bound follows easily from observing that there are only 2^{n^2} possible $(0,1)$-matrices of size $n \times n$. This can be used to show that a primitive semigroup contains a positive matrix which is a product of at most 2^{n^2} copies of the generators. Such a product can be nondeterministically searched for in polynomial space. The lower bound is more involved and uses a reduction from the careful synchronization problem for partial automata [65].

A multiplicative semigroup of nonnegative matrices is called *strongly primitive* if every sufficiently long product of generators is positive. The strong primitivity problem asks whether a given set of matrices generates a strongly primitive semigroup. It is known that if a semigroup of $n \times n$ matrices is strongly primitive, then any product of length at least $2^n - 2$ is positive. This bound is sharp [23]. Thus the strong primitivity problem is also in PSPACE, but the exact complexity is open to the best of our knowledge.

Definition 3.1 (Hurwitz product)**.** The *Hurwitz product* of an ordered set of k matrices associated with a k-tuple $\boldsymbol{m} = (m_1, \ldots, m_k)$ is equal to the sum of all possible products, where for each i the i-th matrix appears exactly m_i times.

A set of nonnegative matrices $\{A_1, \ldots, A_k\}$ is *weakly primitive* (or *k-primitive* or *Hurwitz primitive*) if there exists a tuple $\boldsymbol{m}$ such that the Hurwitz product associated with $\boldsymbol{m}$ is (entrywise) positive. The weak primitivity problem asks whether a given set of matrices is weakly primitive. The complexity of this problem is completely open. However, under the assumption that the matrices have no zero rows, the weak primitivity problem is in PTIME [81].

A similar result holds for the primitivity problem. Under the assumption that the matrices have no zero rows and no zero columns, the primitivity problem for k matrices of size $n \times n$ is decidable in $O(n^2 k)$ time. These results are closely related to the Černý conjecture for synchronizing automata. For a recent survey, see [99].

3.2 Nonnegative powers of a single matrix

3.2.1 Positive Rational Sequences are dominated

Recall that an LRS $\boldsymbol{u}$ is *K-rational* if there exists $d \geq 1$, $c, s \in K^d$ and $M \in K^{d \times d}$ such that $u_n = c^\top M^n s$ for all n.

Proposition 3.1. *An LRS that is both non-degenerate and rational over the semiring $\mathbb{Q}_+$ of nonnegative rational numbers is dominated.*

Proof. Berstel [12] showed that if a sequence $\boldsymbol{u}$ is $\mathbb{Q}_+$-rational then its characteristic roots of maximum modulus all have the form $\rho\omega$ for some non-negative real number ρ and root of unity ω. Since $\boldsymbol{u}$ is non-degenerate it follows that it has a unique dominant root. For an exposition, see [13, Chap. 8, Thm 1.1]. $\square$

We provide a direct proof based on the Perron-Frobenius theorem. We will need the following classical results (see, e.g., [67, Chap. 8]).

Theorem 3.2 (Perron-Frobenius)**.**

1. *If $A \geq 0$ is irreducible then it has* cyclic peripheral spectrum, *i.e., its eigenvalues of maximum modulus have the form $\{\rho, \rho\omega, \ldots, \rho\omega^{k-1}\}$, where $\rho > 0$, k is a positive integer, and ω is a primitive k-th root of unity.*

2. *If a non-negative irreducible matrix A has only one eigenvalue on the spectral circle it is called a primitive matrix. If A is primitive, the pointwise limit* $\lim_{n\to\infty}(A/\rho(A))^n$ *exists and is a strictly positive matrix.*

We now prove Berstel's result for matrix entry recurrences.

Proposition 3.3. *Let $M \in \mathbb{Q}^{d\times d}$ be a non-negative non-degenerate matrix. Then for all $i, j \in \{1, \ldots, d\}$ the LRS $u_n = (M^n)_{i,j}$ is dominated.*

Proof. Since $M \geq 0$ there exists a permutation matrix P such that M can be written in the form $M = PUP^{-1}$, where $U \geq 0$ is block upper triangular. It follows that there exist $i', j' \in \{1, \ldots, d\}$ such that

$$(M^n)_{i,j} = (PU^nP^{-1})_{i,j} = (U^n)_{i',j'}$$

for all $n \in \mathbb{N}$. Now write

$$U = \begin{pmatrix} B_{1,1} & B_{1,2} & \ldots & B_{1,e} \\ 0 & B_{2,2} & \ldots & B_{2,e} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & B_{e,e} \end{pmatrix},$$

where all the blocks in U are non-negative and the diagonal blocks $B_{1,1}, B_{2,2}, \ldots, B_{e,e}$ are irreducible. Then

$$(U^n)_{i',j'} = \sum_{\substack{l_1<l_2<\cdots<l_m \\ n_1+n_2+\cdots n_m=n-(m-1)}} e_{i'}^\top\, B_{l_1,l_1}^{n_1} B_{l_1,l_2} B_{l_2,l_2}^{n_2} \cdots B_{l_{m-1},l_m} B_{l_m,l_m}^{n_m}\, e_{j'}\,, \tag{3.1}$$

where the sum runs over all positive integers m and strictly increasing sequences of block indices $l_1 < \cdots < l_m$ such that i' lies in the l_1 block and j' lies in the l_m block.

Consider a single block $B_{l,l}$ along the diagonal. Since it is irreducible and non-negative, it has cyclic peripheral spectrum. By our assumption of non-degeneracy, $r_l := \rho(B_{l,l}) \geq 0$ is the only eigenvalue on the spectral circle. Thus $B_{l,l}$ is primitive and by our second Perron-Frobenius result above, asymptotically $B^n_{l,l}/r^n_l \sim C_l$ where C_l is a positive matrix and the asymptotic equivalence relation $\sim$ applies entry-wise. Let r_{max} be the maximum spectral radius of a block $B_{l,l}$ lying on a path from i' to j'.

We now analyse the asymptotic behavior of the normalized recurrence $(U^n)_{i',j'}/r^n_{max}$. Consider a summand S_n in $(U^n)_{i',j'}/r^n_{max}$. Replacing the diagonal blocks with their asymptotic limits,

$$S_n \sim \left(\frac{r_{l_1}}{r_{max}}\right)^{n_1} \left(\frac{r_{l_2}}{r_{max}}\right)^{n_2} \cdots \left(\frac{r_{l_m}}{r_{max}}\right)^{n_m} e_{i'}^\top C_{l_1} B_{l_1,l_2} C_{l_2} \cdots B_{l_{m-1},l_m} C_{l_m}\, e_{j'}.$$

Although the number of terms in the sum grows polynomially in n, we see that each term with some $r_{l_k} < r_{max}$ that does not have n_k constant tends to zero exponentially quickly. The remaining summands in $(U^n)_{i',j'}/r^n_{max}$ are thus those where n_k is non-constant only for blocks with $\rho(B_{l_k,l_k}) = r_{max}$. Let K be the sum of the constant powers for non-maximal blocks in such a summand Q_n, and P be the product of the powers of non-maximal r_{l_k}. Then

$$Q_n \sim r^n_{max} \cdot (P/r^{K+m-1}_{max}) \cdot e_{i'}^\top C_{l_1} B_{l_1,l_2} C_{l_2} \cdots B_{l_{m-1},l_m} C_{l_m}\, e_{j'}.$$

Observe that the coefficient of r^n_{max} is a product of spectral radii and non-negative matrices, and is thus non-negative. This implies different such terms containing r^n_{max} cannot cancel out. So $e_{i'}^\top U^n e_{j'} \sim A(n) \cdot r^n_{max}$ for some polynomial A with positive coefficients depending on the non-diagonal blocks and the spectral radii of the diagonal blocks. Thus $(M^n)_{i,j}$ is either dominated by r_{max} or ultimately zero (the latter in case $r_{max} = 0$ or the sum in (3.1) is empty). □

3.2.2 Computing the nonnegative powers of a given matrix

The propositions above allow us to prove the following theorem characterizing the nonnegative powers of a given matrix.

Theorem 3.4. *Given $M \in \mathbb{Q}^{d\times d}$ the set $S := \{n \in \mathbb{N} : M^n \geq 0\}$ is ultimately periodic and effectively computable.*

Proof. Recall that for some (effectively computable) strictly positive integer L the matrix M^L is non-degenerate. It will suffice to show that for each $l \in \{0, \dots, L-1\}$ we can compute the set $S_l := \{n \in S : n \equiv l \bmod L\}$. Our procedure to do this is as follows. First, check for every pair of indices $i, j \in \{1, \dots, d\}$, whether the sequence $(u_n^{(i,j)})_{n=0}^{\infty}$ given by $u_n^{(i,j)} = (M^{Ln+l})_{i,j}$, is dominated. If yes then by Proposition 2.23 we can compute S_l as the intersection over all pairs (i, j) of the sets $\{n \in \mathbb{N} : u_n^{(i,j)} \geq 0\}$. If no, then we claim that S_l is empty.

Indeed, suppose $n_0 \in S_l$. Then for each pair of indices $i, j \in \{1, \dots, d\}$, the LRS $(v_n^{(i,j)})_{n=0}^{\infty}$ defined by $v_n^{(i,j)} = (M^{n_0(1+Ln)})_{i,j} = e_i^\top M^{n_0}(M^{Ln_0})^n e_j^\top$ is both non-degenerate and $\mathbb{Q}^+$-rational. By Proposition 3.1 each sequence $(v_n^{(i,j)})_{n=0}^{\infty}$ is dominated. Moreover, since $(v_n^{(i,j)})_{n=0}^{\infty}$ is a subsequence of $(u_n^{(i,j)})_{n=0}^{\infty}$, the latter is also dominated by Proposition 2.24. This proves (the contrapositive of) our claim. $\square$

Remark. We can extract from the proof of Theorem 3.4 an effective characterisation of those matrices M such that $M^n \geq 0$ for some positive integer n. Let L be the least positive integer such that M^L is non-degenerate. Then some positive power of M is a non-negative matrix iff some positive power of M^L is non-negative iff for all indices (i, j) the sequence $(u_n^{(i,j)})_{n=0}^{\infty}$ defined by $u_n^{(i,j)} := (M^{Ln})_{i,j}$ is dominated and not ultimately negative.

3.3 The Positive Membership Problem for Commutative Semigroups

3.3.1 Eventually Positive Matrices

Recall that a positive matrix is one whose entries are all strictly positive. In this section we show decidability of the following problem.

Problem 3.5 (Positive Membership for Commutative Semigroup). *Given a set of commuting $d \times d$ matrices $\{A_1, \dots, A_k\}$ with rational entries, decide whether the generated semigroup contains a positive matrix.*

We approach this problem through the notion of eventually positive matrix.

Definition 3.2 (Eventually Positive Matrix). We call a matrix M *eventually positive* if there exists a natural number n_0 such that for all $n \geq n_0$, the matrix M^n is positive.

We will need the following definition and result, adapted from Noutsos [72], which characterizes eventual positivity of a matrix by proving a converse of the Perron-Frobenius theorem.

Definition 3.3 (Strong Perron-Frobenius property)**.** A matrix $A \in \mathbb{R}^{n\times n}$ has the *strong Perron-Frobenius property* if there exists an eigenvalue λ with the following properties:

1. λ is real and positive,
2. λ is the unique dominant eigenvalue,
3. λ is simple,
4. λ has a corresponding eigenvector, all of whose entries are positive.

We now have:

Theorem 3.6 (Characterizing eventual positivity [72])**.** *A matrix A is eventually positive iff A and $A^\top$ both have the strong Perron-Frobenius property.*

Clearly a semigroup contains a positive matrix if and only if it contains an eventually positive matrix. By the theorem above, this holds if and only if the semigroup contains a matrix A such that both A and $A^\top$ have the strong Perron-Frobenius property.

3.3.2 Reduction to Integer Programming with Logs of Algebraics

We consider a direct-sum decomposition of $\mathbb{C}^d$ induced by a collection of commuting matrices $A_1, \dots, A_k \in \mathbb{C}^{d\times d}$. Let $\sigma(A_i)$ denote the set of eigenvalues of A_i and consider a tuple of eigenvalues

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

Recall that $\ker(A_i - \lambda_i I)^d$ is the generalized eigenspace of A_i corresponding to λ_i for $i \in \{1, \dots, k\}$. We say that the tuple $\boldsymbol{\lambda}$ is a *joint eigenvalue* of $A_1, \dots, A_k$ if

$$V_{\boldsymbol{\lambda}} := \bigcap_{i=1}^{k} \ker(A_i - \lambda_i I)^d$$

is non null. Intuitively, joint eigenvalues are tuples of eigenvalues of the A_i whose respective generalized eigenspaces have non-trivial intersection. The set of joint eigenvalues is called the *joint spectrum* of $A_1, \dots, A_k$, denoted Σ. Using the fact that commuting matrices preserve each other's generalized eigenspaces, it can be shown (see, e.g., [73, Theorem 2.4]) that

$$\mathbb{C}^d = \oplus_{\boldsymbol{\lambda}\in\Sigma} V_{\boldsymbol{\lambda}}\,,$$

and that, for all $i \in \{1,\dots,k\}$ and $\boldsymbol{\lambda} \in \Sigma$, A_i preserves $V_{\boldsymbol{\lambda}}$, and the restriction of A_i to $V_{\boldsymbol{\lambda}}$ has spectrum $\{\lambda_i\}$. The commuting matrices $A_1^\top,\dots,A_k^\top$ induce an analogous decomposition $\mathbb{C}^d = \oplus_{\boldsymbol{\lambda}\in\Sigma} W_{\boldsymbol{\lambda}}$, where $W_{\boldsymbol{\lambda}} := \cap_{i=1}^k \ker(A_i^\top - \lambda_i I)$.

Consider a matrix $A := A^{m_1}\cdots A_k^{m_k}$ for $m_1,\dots,m_k \in \mathbb{N}$. Given $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_k) \in \Sigma$, by the Spectral Mapping Theorem the restriction of A to the subspace $V_{\boldsymbol{\lambda}}$ has a single eigenvalue $\lambda_1^{m_1}\cdots\lambda_k^{m_k}$. Thus all non-zero vectors in $V_{\boldsymbol{\lambda}}$ are generalized eigenvectors of A for this eigenvalue. It follows that if A is eventually positive, then Conditions 1 and 2 of Theorem 3.6 imply that there exists $\boldsymbol{\lambda} \in \Sigma$ such that $\lambda_1^{m_1}\cdots\lambda_k^{m_k}$ is real positive and is the unique dominant eigenvalue of A. Meanwhile, by Conditions 3 and 4 we can furthermore choose $\boldsymbol{\lambda}$ such that the spaces $V_{\boldsymbol{\lambda}}$ and $W_{\boldsymbol{\lambda}}$ are both one dimensional and contain a positive vector. In such a situation we call $\boldsymbol{\lambda}$ a *dominant joint eigenvalue* for A, $V_{\boldsymbol{\lambda}}$ a *positive right eigenspace*, and $W_{\boldsymbol{\lambda}}$ a *positive left eigenspace.*

Conversely, suppose that there exist $m_1,\dots,m_k \in \mathbb{N}$ and $\boldsymbol{\lambda} \in \Sigma$ such that $\boldsymbol{\lambda}$ is a dominant joint eigenvalue of $A := A_1^{m_1}\cdots A_k^{m_k}$, and $V_{\boldsymbol{\lambda}}$ and $W_{\boldsymbol{\lambda}}$ are positive right and left joint eigenspaces respectively. Then Theorem 3.6 implies that A is eventually positive.

In summary, the semigroup generated by $A_1,\dots,A_k$ contains a positive matrix if and only if there exists $\boldsymbol{\lambda} \in \Sigma$, such that $V_{\boldsymbol{\lambda}}$ and $W_{\boldsymbol{\lambda}}$ are positive joint eigenspaces, and there are $m_1,\dots,m_d \in \mathbb{N}$ with

1. $\prod_{i=1}^k \lambda_i^{m_i} > 0$,
2. $\forall \boldsymbol{\mu} \in \Sigma \setminus \{\boldsymbol{\lambda}\}$, $\prod_{i=1}^k \lambda_i^{m_i} > \left|\prod_{i=1}^k \mu_i^{m_i}\right|$.

It is clear that the Positive Membership Problem reduces to the restricted version of the problem in which one asks for the existence of a positive matrix formed by a product in which all generators of the semigroup are used at least once. Thus we can assume that the desired exponents $m_1,\dots,m_k$ in Conditions 1 and 2 are all strictly positive. In this situation we may rewrite Condition 1 as $\sum_{i=1}^k m_i \arg(\lambda_i) = 0 \bmod 2\pi$ and Condition 2 as either $\mu_1\cdots\mu_k = 0$ or $\sum_{i=1}^k m_i \log|\mu_i/\lambda_i| < 0$ for all $\boldsymbol{\mu} \in \Sigma\setminus\{\boldsymbol{\lambda}\}$. Let $\boldsymbol{c} = (\arg(\lambda_1),\dots,\arg(\lambda_k)) \in \mathbb{R}^k$ and let A be the $(|\Sigma|-1\times k)$-matrix defined by writing, for all $i \in \{1,\dots,k\}$ $\boldsymbol{\mu} \in \Sigma\setminus\{\boldsymbol{\lambda}\}$, $a_{\boldsymbol{\mu},i} := \log|\mu_i/\lambda_i|$ if $\mu_1\cdots\mu_k \neq 0$

and $a_{\boldsymbol{\mu},i} = -1$ if $\mu_1 \cdots \mu_k = 0$. Then the two conditions above are equivalent to the existence of a solution $x \in \mathbb{N}^k$ of the following integer program:

$$(c^\top x = 0 \bmod 2\pi) \wedge Ax < 0 \,.$$

Note that by incorporating positivity constraints $-x_i < 0$ in A we can assume without loss of generality that x ranges over $\mathbb{Z}^k$.

3.3.2.1 Example

Consider the semigroup $\langle A_1, A_2 \rangle$ generated by the matrices:

$$A_1 := \begin{bmatrix} 0.75 & 0.15 & 0.35 & -0.25 \\ 0.15 & 0.75 & -0.25 & 0.35 \\ 0.35 & -0.25 & 0.75 & 0.15 \\ -0.25 & 0.35 & 0.15 & 0.75 \end{bmatrix}, \; A_2 := \begin{bmatrix} 0.75 & 0.3 & 0.2 & -0.25 \\ 0.3 & 0.75 & -0.25 & 0.2 \\ 0.2 & -0.25 & 0.75 & 0.3 \\ -0.25 & 0.2 & 0.3 & 0.75 \end{bmatrix}.$$

On simultaneously diagonalizing A_1 and A_2 we find that

$$A_1 := \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & -0.5 & -0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & -0.5 & -0.5 & 0.5 \end{bmatrix},$$

and

$$A_2 := \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & -0.5 & -0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 \\ 0 & 0 & 1.1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & -0.5 & -0.5 & 0.5 \end{bmatrix}.$$

Consider the joint eigenvalue $(1,1)$. By observing the similarity matrix we can see that the corresponding left and right eigenspaces are spanned by the vector $(1,1,1,1)$, which is positive. Thus the semigroup generated by A_1 and A_2 contains a positive matrix if and only if the $(1,1)$-eigenspace can be made to have a dominant joint eigenvalue.

This occurs if there exist $m_1, m_2 \in \mathbb{N}$ such that

1. $1^{m_1} 1^{m_2} > 0$,
2. (a) $1^{m_1} 1^{m_2} > 1.2^{m_1} 0.9^{m_2}$,
   (b) $1^{m_1} 1^{m_2} > 0.8^{m_1} 1.1^{m_2}$.
   (c) $1^{m_1} 1^{m_2} > 0^{m_1} 0^{m_2}$.

As described in Section 3.3.2, the above conditions can be translated into an integer program (with logs of algebraics). Let $c = [\arg(1), \arg(1)]^\top$ and A be the matrix

$$\begin{bmatrix} \log(1.2) & \log(0.9) \\ \log(0.8) & \log(1.1) \\ -1 & -1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

It is easy to see that the semigroup contains a positive matrix if and only if the integer program (with logs of algebraics)

$$\begin{cases} c^\top x = 0 \bmod 2\pi \\ Ax < 0 \end{cases}$$

is satisfiable.

We now show that the satisfiability of such an integer program is decidable, subject to Schanuel's conjecture.

3.3.3 Integer Programming with Logs of Algebraic Numbers

Problem 3.7 (IP-log)**.** *An instance of the IP-log problem consists of a matrix $A \in \mathbb{R}^{m\times n}$ whose entries are sums of logarithms of non-zero real algebraic numbers and a vector $c \in \mathbb{R}^n$ whose entries are arguments of non-zero algebraic numbers. The problem asks to determine whether there exists a vector $x \in \mathbb{Z}^n$ such that $c^\top x = 0$* mod 2π *and $Ax < 0$.*

Let us start by observing that one can eliminate the equation in the IP-log problem by an effective linear change of variables. Indeed, exponentiation turns the linear relation $c^\top x = 0 \mod 2\pi$ into a multiplicative relation among algebraic numbers. We can then use Theorem 2.5 to find a basis $\{v_1, \dots, v_l\} \subseteq \mathbb{Z}^n$ of the group of integer solutions of the equation $c^\top x = 0 \mod 2\pi$. Let $B \in \mathbb{Z}^{n\times l}$ be the matrix that has these basis vectors as columns. Then $c^\top x = 0 \mod 2\pi$ iff $x = By$ for some integer vector y. Hence the instance of IP-log has a solution iff there exists a vector $y \in \mathbb{Z}^l$ such that $ABy < 0$. In other words, we have reduced the initial instance of IP-log to another instance with a trivial linear constraint.

For an illustration, let $\omega = e^{2\pi i/3}$ be a primitive cube root of unity and let $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ be real positive algebraic numbers. Consider the following system of inequations and an equation:

$$
\begin{aligned}
x_1 \arg(\omega) + x_2 \arg(\omega^2) &= 0 \mod 2\pi \\
x_1 \log \lambda_1 + x_2 \log \lambda_2 &< 0 \\
x_1 \log \lambda_3 + x_2 \log \lambda_4 &< 0
\end{aligned}
$$

Clearly the equation above is satisfied when $x_1 = 3y_1 + 2y_2$ and $x_2 = 3y_1 - y_2$ for some $y_1, y_2 \in \mathbb{Z}$. Thus we eliminate the equation and obtain

$$
\begin{aligned}
(3y_1 + 2y_2) \log \lambda_1 + (3y_1 - y_2) \log \lambda_2 &< 0 \\
(3y_1 + 2y_2) \log \lambda_3 + (3y_1 - y_2) \log \lambda_4 &< 0 \,,
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
y_1 \log \lambda_1^3 \lambda_2^3 + y_2 \log \lambda_1^2/\lambda_2 &< 0, \\
y_1 \log \lambda_3^3 \lambda_4^3 + y_2 \log \lambda_3^2/\lambda_4 &< 0 \,.
\end{aligned}
$$

Theorem 3.8. *The strict homogenous IP-log problem is decidable, assuming Schanuel's conjecture.*

Proof. An instance of the strict homogenous IP-log problem asks to determine the truth of the sentence:

$$\exists x \in \mathbb{Z}^n : (c^\top x = 0 \mod 2\pi) \wedge Ax < 0 \,.$$

As described above, we may assume without loss of generality that the equality constraint is trivial (i.e., $c = 0$). It thus suffices to determine whether the system of inequalities $Ax < 0$ admits a solution $x \in \mathbb{Z}^n$. But, by scaling, this system of inequalities has a solution in $\mathbb{Z}^n$ iff it has a solution in $\mathbb{Q}^n$. In turn, by strictness of the constraints, there is a solution in $\mathbb{Q}^n$ iff there is a solution in $\mathbb{R}^n$. We are thus left with the task of determining whether there exists $x \in \mathbb{R}^n$ such that $Ax < 0$. Here we can apply Fourier-Motzkin elimination [27] (that is to say, quantifier elimination in linear real arithmetic).

Recall that Fourier-Motzkin elimination solves a system of linear inequalities by eliminating the variables sequentially until one obtains an equisatisfiable variable-free system of inequalities between constants. In the case at hand these constants will be rational expressions in a fixed number of logarithms of real algebraic numbers. As such, they are elementary numbers and so, using Richardson's algorithm (Proposition 2.11), we can determine which coefficients are zero. We require Schanuel's conjecture at this point to guarantee termination of the algorithm. For a coefficient that is not zero, we need merely compute it to sufficient precision to determine its sign. □

3.3.3.1 Example, continued

We demonstrate the procedure on the semigroup generated by

$$A_1 := \begin{bmatrix} 0.75 & 0.15 & 0.35 & -0.25 \\ 0.15 & 0.75 & -0.25 & 0.35 \\ 0.35 & -0.25 & 0.75 & 0.15 \\ -0.25 & 0.35 & 0.15 & 0.75 \end{bmatrix}, \ A_2 := \begin{bmatrix} 0.75 & 0.3 & 0.2 & -0.25 \\ 0.3 & 0.75 & -0.25 & 0.2 \\ 0.2 & -0.25 & 0.75 & 0.3 \\ -0.25 & 0.2 & 0.3 & 0.75 \end{bmatrix}.$$

from the previous example. In Section 3.3.2 we performed Step 1, showing that the semigroup contains a positive matrix if and only if the IP-log problem

$$\begin{cases} \begin{bmatrix} \arg(1) & \arg(1) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \bmod 2\pi \\ \begin{bmatrix} \log(1.2) & \log(0.9) \\ \log(0.8) & \log(1.1) \\ -1 & -1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} < \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{cases}$$

is satisfiable.

The equality constraint is trivial and can be dropped. We apply Fourier-Motzkin elimination to the inequalities:

$$\begin{aligned} \log(1.2)x_1 - |\log(0.9)|x_2 &< 0 \\ -|\log(0.8)|x_1 + \log(1.1)x_2 &< 0 \\ -x_1 - x_2 &< 0 \\ -x_1 &< 0 \\ -x_2 &< 0 \end{aligned}$$

Rearranging to isolate x_1:

$$\begin{aligned} &x_1 < \frac{|\log(0.9)|}{\log(1.2)} x_2 \\ &x_1 > \frac{\log(1.1)}{|\log(0.8)|} x_2 \\ &x_1 > -x_2 \\ &x_1 > 0 \\ &x_2 > 0 \end{aligned}$$

After Fourier-Motzkin elimination, we obtain:

$$x_2 > 0$$
$$\max\left(\frac{\log(1.1)}{|\log(0.8)|}x_2, -x_2, 0\right) < x_1 < \frac{|\log(0.9)|}{\log(1.2)}x_2 \,.$$

This system has integer solutions iff $\frac{\log(1.1)}{|\log(0.8)|} - \frac{|\log(0.9)|}{\log(1.2)} < 0$. The truth of the latter inequality can be checked by first using Richardson's algorithm to verify that the left-hand and right-hand expressions are unequal and then calculating to sufficient precision to determine which of the two is greater. In our case, $\frac{\log(1.1)}{|\log(0.8)|} - \frac{|\log(0.9)|}{\log(1.2)} \approx \frac{0.95}{0.223} - \frac{0.105}{0.182} \approx -0.135$. Thus the semigroup generated by A_1 and A_2 contains a positive matrix.

For instance, $(1, 2)$ satisfies the program. The matrix $A_1A_2^2$ is eventually positive and $(A_1A_2^2)^{25}$ is positive.

3.3.4 Algorithm

In summary, we have the following algorithm for the positive membership problem in the commutative case:

INPUT: A set of commuting rational matrices $\{A_1, \ldots, A_k\}$.

1. Find all joint eigenvalues $\boldsymbol{\lambda} \in \Sigma$ for which both the corresponding right eigenspace $V_{\boldsymbol{\lambda}}$ and the left eigenspace $W_{\boldsymbol{\lambda}}$ are one-dimensional and contain a positive vector. If there does not exist a unique such joint eigenvalue, output "NO" and halt.

2. Given unique $\boldsymbol{\lambda}$ identified in Step 1, compute the corresponding IP-log problem as in Section 3.3.2 to see if $\boldsymbol{\lambda}$ is dominant. If the IP-log problem is satisfiable then output "YES" and halt. Otherwise, output "NO" and halt.

We conclude:

Theorem 3.9. *The positive membership problem is decidable for commutative semigroups, assuming Schanuel's conjecture is true.*

Proof. The algorithm above solves the positive membership problem for commutative semigroups, assuming Schanuel's conjecture is true. Correctness follows from the discussion in Section 3.3.2, and the Perron-Frobenius theorem, which implies that a positive matrix can have exactly 1 positive eigenvector. Step 1 may be performed in polynomial time. Unfortunately, the complexity of Step 2 is completely open. This is due to the fact that we have no bound on how closely a nonzero rational function in logs of algebraics may approach zero - such a bound would require a quantitative version of Schanuel's conjecture. □

For practical cases, since near-counterexamples to Schanuel's conjecture are very rare, it might be possible to use a standard computer algebra system or industrial solver to solve the IP-log problem by approximating the constants by rationals and treating it as an integer programming problem.

It would help to use interval arithmetic to ensure that the transcendental constants are approximated to sufficient precision to determine the sign correctly. Alternatively, for the Positive Membership Problem one could verify a proposed integer solution by substituting it back into the eigenvalue product equations, since this merely involves a comparison of algebraic numbers.

3.4 The Non-negative Membership Problem for Commutative Semigroups

We now combine the ideas in Sections 3.2 and 3.3 to solve the problem of deciding whether a semigroup of commuting matrices contains a non-negative matrix. For ease of exposition we will initially assume that the matrices are simultaneously diagonalisable, and then move to the general commuting case.

Problem 3.10 (Non-negative Matrix in Commutative Simultaneously diagonalisable Semigroup)**.** *Given a set of commuting simultaneously diagonalisable $d \times d$ matrices $\{A_1, \dots, A_k\}$ with rational entries, decide whether the semigroup generated by multiplying these matrices together contains a matrix with all its entries greater than or equal to zero.*

First, we refine our notion of dominated recurrences.

Definition 3.4 (Positively dominated by p)**.** Let $\boldsymbol{u} = (u_n)_{n=0}^{\infty}$ be a linear recurrence sequence that is non-degenerate and does not have any polynomial terms in its exponential-polynomial form. Then $\boldsymbol{u} = \sum_{l=1}^{d} c_l \lambda_l^n$ for complex numbers $\lambda_1, \dots, \lambda_d$ and coefficients c_l. We say $\boldsymbol{u}$ is *positively dominated by term p* (here p (for positive) refers to the index) if

1. $c_p > 0$,

2. $\lambda_p > 0$,

3. $\forall l \neq p,\ |\lambda_p| > |\lambda_l|$.

Call this predicate $PD_p(\boldsymbol{u})$.

Given a single matrix M, we consider the d^2 recurrences $\boldsymbol{u}^{ij}$ defined by $(u^{ij})_n := e_i^\top M^n e_j$.

We have shown in Section 3.2 that M being eventually non-negative is equivalent to the decidable condition

$$\forall i \forall j (\boldsymbol{u}^{ij} \text{ is ultimately zero} \vee \exists p\ PD_p(\boldsymbol{u}^{ij})).$$

We now show a similar construction for multiple matrices. Let $A_1, \ldots, A_k \in \mathbb{Q}^{d\times d}$ be a set of commuting simultaneously diagonalisable matrices. The idea is to search for an eventually non-negative matrix $A_1^{m_1} \ldots A_k^{m_k}$.

Let $\boldsymbol{m}$ denote the tuple $(m_1, \ldots, m_k)$. Define the integer parameterized matrix entry recurrence $\boldsymbol{u}^{ij}(\boldsymbol{m})$ by $[\boldsymbol{u}^{ij}(\boldsymbol{m})]_n := e_i^\top [A_1^{m_1} \ldots A_k^{m_k}]^n e_j$.

The existence of an eventually non-negative matrix (and thus, a non-negative matrix) in the semigroup is equivalent to the decidable condition

$$ENN := \exists \boldsymbol{m}\ \forall i \forall j (\boldsymbol{u}^{ij}(\boldsymbol{m}) \text{ is ultimately zero} \vee \exists p\ PD_p(\boldsymbol{u}^{ij}(\boldsymbol{m}))).$$

Let S be a matrix that simultaneously diagonalises the matrices $A_1, \ldots, A_k$ such that $A_r = S^{-1} D_r S = S^{-1} \operatorname{diag}(\lambda_{r1}, \ldots, \lambda_{rd}) S$. The notation λ_{rl} denotes the lth eigenvalue of A_r (with multiplicity).

Then

$$\begin{aligned}
u_n^{ij}(m_1, \ldots, m_k) &:= e_i^\top [A_1^{m_1} \ldots A_k^{m_k}]^n e_j \\
&= e_i^\top S^{-1} [D_1^{m_1} \ldots D_k^{m_k}]^n S e_j \\
&= e_i^\top S^{-1} [\operatorname{diag}(\lambda_{11}, \ldots, \lambda_{1d})^{m_1} \ldots \operatorname{diag}(\lambda_{k1}, \ldots, \lambda_{kd})^{m_k}]^n S e_j \\
&= e_i^\top S^{-1} \left[\operatorname{diag}\left(\prod_{r=1}^{k} \lambda_{r1}^{m_r}, \ldots, \prod_{r=1}^{k} \lambda_{rd}^{m_r}\right)\right]^n S e_j \\
&= \sum_{l=1}^{d} (S^{-1})_{il}\, (S)_{lj} \left[\prod_{r=1}^{k} \lambda_{rl}^{m_r}\right]^n .
\end{aligned}$$

Now we have an exponential-polynomial representation for $\boldsymbol{u}^{ij}(\boldsymbol{m})$ with coefficients $(S^{-1})_{il}(S)_{lj}$ and roots $\prod_{r=1}^{k} \lambda_{r1}^{m_r}, \ldots, \prod_{r=1}^{k} \lambda_{rd}^{m_r}$.

We see that $\boldsymbol{u}^{ij}(\boldsymbol{m})$ is positively dominated by p if

1. $(S^{-1})_{ip}\, (S)_{pj} > 0$,

2. $\prod_{r=1}^{k} \lambda_{rp}^{m_r} > 0$,

3. $\forall l \neq p$ such that $(S^{-1})_{il}\,(S)_{lj} \neq 0$, $\left|\prod_{r=1}^{k} \lambda_{rp}^{m_r}\right| > \left|\prod_{r=1}^{k} \lambda_{rl}^{m_r}\right|$.

Let $c(p) = (\arg(\lambda_{1p}), \ldots, \arg(\lambda_{kp}))$ and let $A(p)$ be the (at most) $(d-1)\times k$-matrix defined by $a_{lr} = \log|\lambda_{rl}/\lambda_{rp}|$. Here l runs over elements of $\{1, \ldots, d\}$ apart from p such that $(S^{-1})_{il}\,(S)_{lj} \neq 0$. Let $\boldsymbol{m}$ be the vector $(m_1, \ldots, m_k) \in \mathbb{N}^k$.

Now Condition 2 is equivalent to $c(p)^\top \boldsymbol{m} = 0 \bmod 2\pi$ and Condition 3 is equivalent to $A(p)\boldsymbol{m} < 0$.

For simplicity we introduce the following new notation:

$$\begin{aligned}
Z(i,j,\boldsymbol{m}) &:= \boldsymbol{u}^{ij}(\boldsymbol{m}) \text{ is identically zero}\\
S(i,j,p) &:= (S^{-1})_{ip}\,(S)_{pj} > 0\\
C(p,\boldsymbol{m}) &:= c(p)^\top \boldsymbol{m} = 0 \bmod 2\pi\\
A(i,j,p,\boldsymbol{m}) &:= A(p)\boldsymbol{m} < 0\\
NND(i,j,p,m) &:= Z(i,j,\boldsymbol{m}) \vee (S(i,j,p) \wedge C(p,\boldsymbol{m}) \wedge A(i,j,p,\boldsymbol{m})).
\end{aligned}$$

Note that $A(i,j,p,\boldsymbol{m})$ depends on i and j because we only care about the eigenvalue blocks where the coefficient is non-zero.

Writing out the predicate ENN in full with new notation, we have

$$\begin{aligned}
ENN &:= \exists \boldsymbol{m}\ \forall i \forall j (\boldsymbol{u}^{ij}(\boldsymbol{m}) \text{ is identically zero} \vee \exists p\ PD_p(\boldsymbol{u}^{ij}(\boldsymbol{m})))\\
&\equiv \exists \boldsymbol{m}\ \forall i \forall j (Z(i,j,\boldsymbol{m}) \vee \exists p\ (S(i,j,p) \wedge C(p,\boldsymbol{m}) \wedge A(i,j,p,\boldsymbol{m})))\\
&\equiv \exists \boldsymbol{m}\ \forall i \forall j\ \exists p\ NND(i,j,p,\boldsymbol{m}).
\end{aligned}$$

Now observe that the quantifications over i, j and p are finite and range over $\{1, \ldots, d\}$. Thus we can replace the quantifications with finite disjunctions and conjunctions.

Our goal is to move all disjunctions outside the existential quantifier on $\boldsymbol{m}$ so that we can use the IP-log algorithm from Section 3.3.3 on conjunctions of terms of the form $A\boldsymbol{m} < 0$.

Let f range over functions assigning a particular choice of p to each sequence $\boldsymbol{u}^{ij}(\boldsymbol{m})$. There are $d^{(d\times d)}$ such functions.

$$\begin{aligned}
ENN &\equiv \exists \boldsymbol{m}\ \forall i \forall j (\exists p\ NND(i,j,p,\boldsymbol{m}))\\
&\equiv \exists \boldsymbol{m}\ \wedge_{ij}\ (\vee_p\ NND(i,j,p,\boldsymbol{m}))\\
&\equiv \exists \boldsymbol{m}\ \vee_f\ (\wedge_{ij} NND(i,j,p = f(i,j),\boldsymbol{m}))\\
&\equiv \vee_f(\ \exists \boldsymbol{m}\ \wedge_{ij}\ NND(i,j,p = f(i,j),\boldsymbol{m})).
\end{aligned}$$

Essentially this means we non-deterministically choose a p for each sequence $\boldsymbol{u}^{ij}(\boldsymbol{m})$ and then check if there is an $\boldsymbol{m}$ that works for all of them — that makes all the selected p's into real positively dominating terms. Analysing the elements of $NND(i,j,f(i,j),\boldsymbol{m})$, we see that $Z(i,j,\boldsymbol{m})$ and $S(i,j,f(i,j))$ constraints are trivially checkable. Constraints $\wedge_{ij}C(f(i,j),\boldsymbol{m})$ can be removed iteratively using Masser's theorem. The remaining conjunctions are terms of the form $\wedge_{ij}A(i,j,f(i,j),\boldsymbol{m})$, but since these are linear programs $A(i,j,p)\boldsymbol{m} < \boldsymbol{0}$ we can simply concatenate the various matrices $A(i,j,p)$ together. We can then use the IP-log algorithm from Section 3.3.3 to check if there exists an integer solution to the conjunction of these terms.

Of course in practice we would only need to check d different matrices of size $d-1 \times k$ since the eigenvalues are the same for all sequences.

Thus we have reduced the diagonalisable case to a finite disjunction of predicates, each of which can be reduced to strict homogenous IP-log.

More generally, we consider:

Problem 3.11 (Non-negative Membership for Commutative Semigroup)**.** *Given a set of commuting $d \times d$ matrices $\{A_1, \ldots, A_k\}$ with rational entries, decide whether the semigroup generated by multiplying these matrices together contains a matrix with all its entries greater than or equal to zero.*

It is known that unfortunately simultaneous Jordanization of commuting matrices is not always possible [18]. However, a slightly weaker block diagonal form [73, Thm 12] is possible. Here we put the matrices into block diagonal form, where each block is of the form $\lambda_i I + N$ where N is strictly upper triangular and thus nilpotent.

Let $A_1, \ldots, A_k \in \mathbb{Q}^{d\times d}$ be a set of commuting matrices.

As before, define the integer parameterized matrix entry recurrence $\boldsymbol{u}^{ij}(\boldsymbol{m})$ by $u_n^{ij}(m_1,\ldots,m_k) := e_i^\top [A_1^{m_1} \ldots A_k^{m_k}]^n e_j$.

Let S be a matrix that simultaneously block-diagonalises the matrices such that $A_r = S^{-1}BD_rS = S^{-1}\operatorname{diag}(B_{r1},\ldots,B_{rb})S$. Here $B_{rl} = \lambda_{rl}I + N_{rl}$ denotes the lth block of A_r, where N_{rl} is strictly upper triangular. Note that for fixed l the various

N_{rl} inherit commutativity from the original matrices. Then we have that

$$\begin{aligned}
u_n^{ij}(m_1,\ldots,m_k) &:= e_i^\top [A_1^{m_1}\ldots A_k^{m_k}]^n e_j \\
&= e_i^\top S^{-1}[BD_1^{m_1}\ldots BD_k^{m_k}]^n S e_j \\
&= e_i^\top S^{-1}[\mathrm{diag}(B_{11},\ldots,B_{1b})^{m_1}\ldots \mathrm{diag}(B_{k1},\ldots,B_{kb})^{m_k}]^n S e_j \\
&= e_i^\top S^{-1}\left[\mathrm{diag}\left(\prod_{r=1}^k B_{r1}^{m_r},\ldots,\prod_{r=1}^k B_{rb}^{m_r}\right)\right]^n S e_j \\
&= [s_{i1}^{-1},\ldots,s_{id}^{-1}]\,\mathrm{diag}\left(\left[\prod_{r=1}^k B_{r1}^{m_r}\right]^n,\ldots,\left[\prod_{r=1}^k B_{rb}^{m_r}\right]^n\right)^n [s_{1j},\ldots,s_{dj}]^\top.
\end{aligned}$$

Let us examine the structure of the submatrix $\left[\prod_{r=1}^k B_{rb}^{m_r}\right]^n$. Recall that $B_{rl} = \lambda_{rl} I + N_{rl}$. Note that any power or product of powers of the nilpotents can be non-zero only upto total degree at most d. We adopt the convention that a zero power indicates the identity matrix of appropriate size. For ease of notation we drop the block subscript and expand out

$$\begin{aligned}
\left[\prod_{r=1}^k B_{rb}^{m_r}\right]^n &= \left[\prod_{r=1}^k (\lambda_r I + N_r)^{m_r}\right]^n \\
&= \prod_{r=1}^k \lambda_r^{nm_r}(I + N_r/\lambda_r)^{nm_r} \\
&= \prod_{r=1}^k \lambda_r^{nm_r}\cdot \prod_{r=1}^k\left(\sum_{i_r=0}^d \binom{nm_r}{i_r}(N_r/\lambda_r)^{i_r}\right) \\
&= \prod_{r=1}^k \lambda_r^{nm_r}\cdot\left[\sum_{(i_1,\ldots,i_k)=(0,\ldots,0)}^{i_1+\cdots+i_k=d}\left(\prod_{r=1}^k\binom{nm_r}{i_r}(N_r/\lambda_r)^{i_r}\right)\right] \\
&= \mathrm{MPoly}(n\mathbf{m})\prod_{r=1}^k \lambda_r^{nm_r}.
\end{aligned}$$

Here $\mathrm{MPoly}(n\mathbf{m})$ is shorthand for the block-matrix with entries that are polynomial of degree at most d in the variables $(m_1,\ldots,m_k)$ (scaled by n) with coefficients arising from the nilpotent submatrices.

Substituting this back into our original expression for $u_n^{ij}(m_1,\ldots,m_k)$ we get

$$
\begin{aligned}
&u_n^{ij}(m_1,\ldots,m_k) := e_i^\top [A_1^{m_1}\ldots A_k^{m_k}]^n e_j\\
&= [s_{i1}^{-1},\ldots,s_{id}^{-1}]\,\mathrm{diag}\left(\left[\prod_{r=1}^{k} B_{r1}^{m_r}\right]^n,\ldots,\left[\prod_{r=1}^{k} B_{rb}^{m_r}\right]^n\right)^n [s_{1j},\ldots,s_{dj}]^\top\\
&= [s_{i1}^{-1},\ldots,s_{id}^{-1}]\,\mathrm{diag}\left(\mathrm{MPoly}_1(n\mathbf{m})\prod_{r=1}^{k}\lambda_{r1}^{nm_r},\ldots,\mathrm{MPoly}_b(n\mathbf{m_r})\prod_{r=1}^{k}\lambda_{rd}^{nm_r}\right)^n [s_{1j},\ldots,s_{dj}]^\top\\
&= \sum_{l=1}^{d}\left(\mathrm{poly}_l^{ij}(n\mathbf{m})\prod_{r=1}^{k}\lambda_{rl}^{nm_r}\right) \text{(after folding in constants).}
\end{aligned}
$$

Notice that the asymptotic top term in n in the polynomial $\mathrm{poly}_l^{ij}(n\mathbf{m})$ is the homogenous subpolynomial of highest degree in $(m_1,\ldots,m_k)$ - call it $h_l^{ij}(\mathbf{m})$. Once we pick a particular p to be our positively dominating term, we only need the following three conditions for the recurrence to be positively dominated by p:

1. $h_p^{ij}(\mathbf{m}) > 0$,
2. $\prod_{r=1}^{k}\lambda_{rp}^{m_r} > 0$,
3. $\forall l \neq p$ such that $\mathrm{poly}_l^{ij}(n\mathbf{m})$ is not the zero polynomial, $\left|\prod_{r=1}^{k}\lambda_{rp}^{m_r}\right| > \left|\prod_{r=1}^{k}\lambda_{rl}^{m_r}\right|$.

Via the same algebraic manipulations as in the diagonalisable case, checking

$$ENN := \exists m\ \forall i \forall j(\boldsymbol{u}^{ij}(\boldsymbol{m}) \text{ is identically zero} \vee \exists p\ PD_p(\boldsymbol{u}^{ij}(\boldsymbol{m}))),$$

reduces to solving conjunctions of the form

$$\wedge_{ij}(h_l^{ij}(\mathbf{m}) > 0 \wedge \mathbf{c}(p)^\top\mathbf{m} = 0 \bmod 2\pi \wedge \mathbf{A}(p)\mathbf{m} < 0)$$

over the integers.

We can eliminate the second conjunct using Masser's theorem. Now suppose there exists a real solution on the unit sphere to $h_l^{ij}(\mathbf{m}) > 0 \wedge \mathbf{A}(p)\mathbf{m} < 0$. By openness, there exists a rational solution nearby. Since both these conjuncts are homogenous, $\mathbf{m}$ is a solution iff $n\mathbf{m}$ is a solution, for all real $n > 0$. Thus we may clear denominators from the rational solution to obtain an integer solution $\mathbf{m}$ to ENN.

The sentence

$$\exists\mathbf{m}\in\mathbb{R}^k : h_l^{ij}(\mathbf{m}) > 0 \wedge \mathbf{A}(p)\mathbf{m} < 0$$

can be written in the first order theory of the reals with exponentiation, which is decidable assuming Schanuel's conjecture as shown by Wilkie and Macintyre [63].

We now need to prove that failing to find such a real solution implies that the semigroup does not contain a non-negative matrix. Suppose that there exists some $m = (m_1, \dots, m_k)$ such that $A_1^{m_1} \dots A_k^{m_k}$ is non-negative. Then $(A_1^{m_1} \dots A_k^{m_k})^n$ is non-negative for all n. By Proposition 3.1, each individual recurrence is ultimately zero or must have a strictly dominant top term. Thus m satisfies $Am < 0$ and $c^\top m = 0 \bmod 2\pi$ for the appropriate A and c. The top term (as a function of n) has a polynomial coefficient which is the homogenous polynomial we identified above. Thus the homogenous polynomial is non-negative for m, completing the requirements necessary for the sentence $\exists \mathbf{m} \in \mathbb{R}^k : h_l^{ij}(\mathbf{m}) > 0 \wedge \mathbf{A}(p)\mathbf{m} < 0$ to be true.

3.4.1 Algorithm

In summary, we have the following algorithm for the non-negative membership problem in the commutative case:

INPUT: A set of commuting $d \times d$ rational matrices $\{A_1, \dots, A_k\}$.

1. Simultaneously block-diagonalise the matrices to get d^2 integer-parameterized recurrences $\boldsymbol{u}^{ij}(\boldsymbol{m})$.

2. Construct the predicate $ENN := \exists m\ \forall i \forall j (\boldsymbol{u}^{ij}(\boldsymbol{m})$ is identically zero$\vee \exists p\ PD_p(\boldsymbol{u}^{ij}(\boldsymbol{m})))$ and reduce it to exponentially many disjuncts of the form $\exists \mathbf{m} \in \mathbb{R}^k : h_l^{ij}(\mathbf{m}) > 0 \wedge \mathbf{A}(p)\mathbf{m} < 0$

3. Solve each disjunct using the algorithm of Wilkie and Macintyre for the decidability of the real exponential field [63].

We conclude:

Theorem 3.12. *The non-negative membership problem is decidable for commutative semigroups, assuming Schanuel's conjecture is true.*

Proof. The algorithm above solves the non-negative membership problem for commutative semigroups, assuming Schanuel's conjecture is true. Correctness follows from the extended discussion in Section 3.4. Step 1 can be performed in polynomial time. Assuming an oracle for Step 3, Step 2 can be performed in nondeterministic polynomial time, by guessing the correct choice of p for each of the d^2 recurrences. Unfortunately Step 3 itself has no known complexity bound, due to the use of Schanuel's conjecture. □

3.5 Undecidability in the Non-commutative Case

To complement the decidability results for commuting matrices in the preceding sections, in this section we show undecidability of the full version of the membership problem, in which commutativity is not assumed:

Problem 3.13 (Non-negative Membership)**.** *Given a set of k rational $d \times d$ matrices, decide whether the semigroup they generate contains a non-negative matrix.*

The proof of undecidability is by reduction from the Halfspace Reachability Problem.

Problem 3.14 (Halfspace Reachability Problem with scalar 1/2)**.** *Given vectors u and v in $\mathbb{Q}^d$ and a matrix semigroup $\mathcal{S}$ generated by matrices $\{A_1, \ldots, A_k\} \in \mathbb{Q}^{d \times d}$, decide whether there exists a matrix $A \in \mathcal{S}$ such that $u^\top A v \geq 1/2$.*

By Theorem 2.25, this problem is undecidable for 5 matrices in dimension 10.

Theorem 3.15. *The Non-negative Membership Problem is undecidable for 7 matrices in dimension 12.*

Proof. Given non-negative integers m, n, write $0^{m \times n}$ for the zero matrix of dimension $m \times n$. Suppose that we are given an instance of the halfspace reachability problem, defined by vectors $u, v \in \mathbb{Q}^d$ and a matrix semigroup $\mathcal{S}$ generated by matrices $A_1, \ldots, A_k \in \mathbb{Q}^{d \times d}$. Now consider the semigroup $\mathcal{S}'$ generated by the following matrices of dimension $(d+2) \times (d+2)$:

$$U := \begin{bmatrix} 1 & -1/2 & \boldsymbol{u}^\top \\ 0 & 0 & 0^{1 \times d} \\ 0^{d \times 1} & 0^{d \times 1} & 0^{d \times d} \end{bmatrix},$$

$$A_i' := \begin{bmatrix} 1 & -1/2 & 0^{1 \times d} \\ 0 & 0 & 0^{1 \times d} \\ 0^{d \times 1} & 0^{d \times 1} & A_i \end{bmatrix} \qquad (i = 1, \ldots, k),$$

$$\text{and } V := \begin{bmatrix} 1 & -1/2 & 0^{1 \times d} \\ 0 & 0 & 0^{1 \times d} \\ 0^{d \times 1} & \boldsymbol{v} & 0^{d \times d} \end{bmatrix}.$$

Note that matrix A_i' incorporates A_i for $i = 1, \ldots, k$, while the matrices U and V respectively incorporate the initial and final vectors u and v of the probabilistic automaton.

We claim that the semigroup $\mathcal{S}'$ contains a non-negative matrix if and only if there exists a matrix $A \in \mathcal{S}$ such that $u^\top A v \geq \frac{1}{2}$. To this end, consider a string of matrices chosen from the set $\{U, V\} \cup \{A_1', \ldots, A_k'\}$.

We proceed via a case analysis:

1. Strings not containing both a U and a V. By direct multiplication, one can observe that the product B of any such string has $B_{1,2} = -\frac{1}{2}$ and hence cannot be a non-negative matrix.

2. Strings containing both a U and a V. Observe by direct multiplication that $UU = A'_i U = VU = U$. Thus the product of any such string is equal to the product of its suffix starting from the *last* occurrence of U.

   (a) Suffixes not containing a V. See Case 1.

   (b) Suffixes not ending in their *first* occurrence of V. Define N to be the $(d+2) \times (d+2)$ matrix such that $N_{1,1} = 1, N_{1,2} = -\frac{1}{2}$, and $N_{i,j} = 0$ for all other entries.

   Observe that $VV = VA'_i = N$. Also $UN = A'_i N = VN = NN = N$. Thus any product from a suffix that does not end with its *first* occurrence of V is N, which is not a non-negative matrix.

   (c) Suffixes of the form $UA'_{i_1} \cdots A'_{i_s} V$, for some $i_1, \ldots, i_s \in \{1, \ldots, k\}$. In this case we have

   $$B_{1,2} = (UA'_{i_1} \cdots A'_{i_s} V)_{1,2} = u^\top A_{i_1} \cdots A_{i_s} v - \frac{1}{2}.$$

   Since it further holds that $B_{1,1} = 1$ and $B_{i,j} = 0$ for all other entries (i, j), B is non-negative if and only if $u^\top A_{i_1} \cdots A_{i_s} v \geq \frac{1}{2}$.

We conclude that there exists a non-negative matrix in the semigroup $\mathcal{S}'$ if and only if there exists a matrix $A \in \mathcal{S}$ such that $u^\top Av \geq \frac{1}{2}$.

Since we need two additional dimensions and two additional matrices to implement our reduction, there is an undecidable instance of Non-negative Membership for 7 matrices in dimension 12. □

3.5.1 The Positive Membership Problem

Problem 3.16 (Positive Membership). *Given a set of k rational $d \times d$ matrices, decide whether the generated semigroup contains a positive matrix.*

For arbitrary rational noncommuting matrices, we conjecture that the Positive Membership Problem is undecidable, but we have not been able to prove this. It forms a natural open problem for future research.

We now discuss a special case.

Although nearly all decision problems involving noncommuting matrices are undecidable, there is a notable exception in the field of quantum automata:

Problem 3.17 (Strict Threshold Problem for Quantum Automata)**.** *Given a threshold $\lambda \in \mathbb{Q}$, a vector $s \in \mathbb{Q}^d$ and a projection matrix $P \in \mathbb{Q}^{d\times d}$, and a matrix semigroup $\mathcal{S}$ generated by unitary matrices $\{U_1, \ldots, U_k\} \in \mathbb{Q}^{d\times d}$, decide whether there exists a matrix $U \in \mathcal{S}$ such that $\left\|s^\top UP\right\|_2 > \lambda$.*

This problem is decidable, although the non-strict version is undecidable [16]. The proof relies centrally on the fact that a compact subsemigroup of a topological group of real matrices has an algebraic description [95]. However this result applies only to diagonalizable unimodular matrices (all eigenvalues of modulus 1) such as unitary matrices.

For such matrices, the Positive Membership Problem is trivially decidable.

Theorem 3.18. *No unimodular matrix in dimension ≥ 2 is positive.*

Proof. By the Perron-Frobenius theorem, a positive matrix in 2 or more dimensions has a unique eigenvalue which is simple and strictly larger than all other eigenvalues. A unimodular matrix has all eigenvalues of modulus 1, so it cannot be positive. □

3.6 Conclusion

We can summarise the current state of knowledge on the positive and non-negative membership problems for matrix semigroups with rational entries in the following table:

Table 3.1: Does a ... rational matrix semigroup contain a ... ?

generator type	**positive matrix**	**non-negative matrix**
single	Decidable [72] (Noutsos 2006)	Decidable
commuting	Decidable assuming Schanuel (via Richardson)	Decidable assuming Schanuel (via Macintyre & Wilkie)
noncommuting	Open (conjecture undecidable)	Undecidable

We leave open the question of quantitative refinements of our decidability results. These include giving complexity upper bounds for the Non-negative and Positive Membership Problems as well as the related question of giving upper bounds on the length of the shortest string of generators that yields a non-negative or positive matrix in a given semigroup. Both questions would seem to be difficult owing to the use of Schanuel's conjecture in our proofs.

Characterising the complexity of determining whether a matrix is eventually non-negative would seem to be more straightforward. We claim that the decision procedure can be implemented in non-deterministic polynomial time. On the other hand, our efforts at proving an NP-hardness lower bound have been stymied by the strong closure properties of the set of non-negative matrices. Could it be that the problem is in PTIME? Note that the analogous Eventual Positivity problem is in PTIME [72] and so is the Primitivity problem for one matrix. However it is possible to construct NP-hard instances of the Skolem problem using a $(0,1)$-matrix [85].

Another direction in which our results could be extended is the case of different rings. While we only consider the rationals, the analogous problems over finite fields would clearly be unconditionally decidable since the semigroups are finite. The complexity of such problems would be interesting to explore, with results possibly analogous to the Primitivity problems discussed in Section 3.1.1

Chapter 4

Pseudo-orbits and the Pseudo-Skolem Problem

4.1 Introduction

The orbit and Skolem problems are defined on the exact dynamics of the linear system. In dynamical systems theory, one is often interested in "rough" dynamics of a system—in topological terms, we wish to study closed sets containing the orbit. Orbits arising from linear dynamics are usually not closed sets. Indeed the orbit of the point 1 under the map $x \mapsto \frac{1}{2}x$ does not contain the limit point 0. One way to retain closure is through *pseudo-orbits* [26], a concept going back several decades. A pseudo-orbit generalizes the orbit by allowing arbitrarily small imprecisions throughout the dynamics.

Let (A, s) be a linear dynamical system. For a precision $\epsilon > 0$, we say t is in the *ϵ-pseudo-orbit* of s if there is a sequence of points $(s = s_0, s_1, \ldots, s_n = t)$ with $n > 0$ such that $\|As_i - s_{i+1}\| < \epsilon$ for each $i \in \{0, \ldots, n-1\}$. That is, an ϵ-pseudo-orbit contains a sequence of points that would be indistinguishable from an orbit if each state were known only up to precision ϵ. Finally, t is in the *pseudo-orbit* of s if it is in the ϵ-pseudo-orbit of s *for all* $\epsilon > 0$.

One can provide a computational analogue of pseudo-orbits (see [79]). Alice is simulating the trajectory of a dynamical system, but in every iteration her computation has a rounding error ϵ. An infinitely powerful adversary, Bob, rounds Alice's result in an arbitrary fashion to a new state within a distance of ϵ of the actual outcome. A state t is pseudo-reachable from s iff Bob can fool Alice into believing that t is reachable in the simulation no matter how accurate her simulation is.

We can formulate analogous decision problems on pseudo-orbits. The *pseudo-orbit problem* asks, given a linear dynamical system and two states s and t, whether

t is in the pseudo-orbit of s. The *hyperplane pseudo-reachability* (or pseudo-Skolem) problem asks, given a linear dynamical system, an initial state s, and a hyperplane, if there is an ϵ-pseudo orbit from s that intersects the hyperplane for every $\epsilon > 0$.

In this chapter, we study reachability problems for pseudo-orbits of linear dynamical systems. We show that the pseudo-orbit problem is decidable in polynomial time and that the Skolem problem is decidable in full generality on pseudo-orbits.

First, we generalize Kannan and Lipton's analysis to show that the pseudo-orbit problem can be decided in polynomial time. Our proof involves a careful examination of the eigenvalues of the matrix A, similar to Kannan and Lipton's proof. More generally, we show that pseudo-reachability to a bounded semialgebraic set is decidable.

Moving to unbounded target sets, we consider the hyperplane pseudo-reachability (a.k.a. pseudo-Skolem) problem. Our proof again proceeds by a case analysis on the eigenvalues of A. The most challenging case is when there is an eigenvalue of modulus greater than 1. We analyze a series whose terms are polynomial-exponential functions of $n \in \mathbb{N}$ associated with the dynamics. We show that the infimum of this sum can be effectively computed. The proof of effective computability uses tools from Diophantine approximation as well as a reduction to the decision problem for the theory of real closed fields.

We show that the dynamics pseudo-reaches the hyperplane in case the infimum of the above sum is 0. If the infimum is non-zero, we prove that we can find an effective bound N such that the dynamics pseudo-reaches the hyperplane iff, for sufficiently small ϵ, it pseudo-reaches the hyperplane within N steps.

Putting everything together, we conclude that the pseudo-Skolem problem is decidable.

4.2 Pseudo-orbits

An m-dimensional discrete-time linear dynamical system is specified by an $m \times m$ matrix A of rational numbers. The *trajectory* determined by an initial state $x_0 \in \mathbb{R}^m$ is the sequence $(x_n)_{n\geq 0}$ given by

$$x_{n+1} = Ax_n, \quad (n \in \mathbb{N}).$$

We call the set $\mathcal{O}(A, x_0) := \{x_n \mid n \in \mathbb{N}\}$ the *orbit* of x_0.

For any $\epsilon > 0$, an ϵ-perturbed linear dynamical system has state trajectories $(x_n)_{n\geq 0}$ such that

$$x_{n+1} = Ax_n + d_n, \quad (n \in \mathbb{N}),$$

where A is as before and $d_n \in [-\epsilon, \epsilon]^m$ for all n. For an initial state $x_0 \in \mathbb{R}^m$, we define the *ϵ-pseudo-orbit* $\widetilde{\mathcal{O}}_\epsilon(A, x_0)$ of the dynamics as the set of states reachable in the perturbed dynamics. More formally, define

- for $n = 0$, $\widetilde{\mathcal{O}}_\epsilon^{(n)}(A, x_0) := \{x_0\}$,
- for all $n \in \mathbb{N}$, $\widetilde{\mathcal{O}}_\epsilon^{(n+1)}(A, x_0) := \{Ax + d \in \mathbb{R}^m \mid x \in \widetilde{\mathcal{O}}_\epsilon^{(n)}(A, x_0), d \in [-\epsilon, \epsilon]^m\}$, and
- $\widetilde{\mathcal{O}}_\epsilon(A, x_0) := \bigcup_{n\geq 0} \widetilde{\mathcal{O}}_\epsilon^{(n)}(A, x_0)$.

Finally, we define the *pseudo-orbit* $\widetilde{\mathcal{O}}(A, x_0) := \bigcap_{\epsilon>0} \widetilde{\mathcal{O}}_\epsilon(A, x_0)$ as the intersection of all the ϵ-pseudo orbits of x_0, for all $\epsilon > 0$. Clearly, $\mathcal{O}(A, x) \subseteq \widetilde{\mathcal{O}}(A, x)$ for any A and x.

We will make use of the following characterization, which follows directly from the definition: Any $t \in \widetilde{\mathcal{O}}_\epsilon(A, s)$ is of the form $t = A^n s + \sum_{i=0}^{n-1} A^i d_{n-i-1}$ for some $n \in \mathbb{N}$ and some sequence of perturbations d_i with $\|d_i\|_\infty \leq \epsilon$.

We also need the following properties of $\widetilde{\mathcal{O}}_\epsilon(A, x)$ and $\widetilde{\mathcal{O}}(A, x)$.

Claim 4.1 (Transitivity)**.** *For every A and $\epsilon > 0$, and for states $s, t, u \in \mathbb{R}^m$, if $t \in \widetilde{\mathcal{O}}_\epsilon(A, s)$ and $u \in \widetilde{\mathcal{O}}_\epsilon(A, t)$ then $u \in \widetilde{\mathcal{O}}_\epsilon(A, s)$. If $t \in \widetilde{\mathcal{O}}(A, s)$ and $u \in \widetilde{\mathcal{O}}(A, t)$, then $u \in \widetilde{\mathcal{O}}(A, s)$.*

Claim 4.2 (Closure)**.** *For every A, $\epsilon > 0$, and state $s \in \mathbb{R}^m$, the sets $\widetilde{\mathcal{O}}_\epsilon(A, s)$ and $\widetilde{\mathcal{O}}(A, s)$ are closed sets.*

Problem 4.3 (Orbit problem)**.** *Given $A \in \mathbb{Q}^{m\times m}$ and $s, t \in \mathbb{Q}^m$, decide whether $t \in \mathcal{O}(A, s)$.*

A celebrated result of Kannan and Lipton [53] shows that the Orbit Problem is decidable in polynomial time.

Theorem 4.4. *[53] The orbit problem is decidable in polynomial time.*

4.3 The pseudo-orbit problem is decidable in polynomial time

In this section, we study the pseudo-orbit problem.

Problem 4.5 (Pseudo-orbit problem)**.** *Given* $A \in \mathbb{Q}^{m\times m}$ *and* $s, t \in \mathbb{Q}^m$, *decide whether* $t \in \widetilde{\mathcal{O}}(A, s)$.

We will show that Problem 4.5 is decidable in polynomial time, by reducing to the orbit problem.

First we establish that pseudo-orbits can be translated with change of bases.

Proposition 4.6. *For matrices* $A, B, Q \in \mathbb{R}^{m\times m}$ *with* $A = QBQ^{-1}$ *and for any* $x \in \mathbb{R}^m$, *we have* $Q\,\widetilde{\mathcal{O}}_{\gamma_2}(B, Q^{-1}x) \subseteq \widetilde{\mathcal{O}}_{\epsilon}(A, x) \subseteq Q\,\widetilde{\mathcal{O}}_{\gamma_1}(B, Q^{-1}x)$, *where* $\gamma_1 = \epsilon \left\|Q^{-1}\right\|_\infty$ *and* $\gamma_2 = \epsilon / \left\|Q\right\|_\infty$. *Moreover,* $\widetilde{\mathcal{O}}(A, x) = Q\,\widetilde{\mathcal{O}}(B, Q^{-1}x)$.

Proof. We want to show

$$Q\,\widetilde{\mathcal{O}}_{\gamma_2}(B, Q^{-1}x) \subseteq \widetilde{\mathcal{O}}_{\epsilon}(A, x) \subseteq Q\,\widetilde{\mathcal{O}}_{\gamma_1}(B, Q^{-1}x), \tag{4.1}$$

where $\gamma_1 = \epsilon \left\|Q^{-1}\right\|_\infty$ and $\gamma_2 = \epsilon / \left\|Q\right\|_\infty$.

Take any $y \in \widetilde{\mathcal{O}}_{\epsilon}(A, x)$. We show that $Q^{-1}y \in \widetilde{\mathcal{O}}_{\gamma_1}(B, Q^{-1}x)$ to get the right-hand side of (4.1). Since $y \in \widetilde{\mathcal{O}}_{\epsilon}(A, x)$, there is a state trajectory $(x_0, x_1, \ldots)$ and a sequence $(d_0, d_1, \ldots)$ such that $x_0 = x$, $x_{n+1} = Ax_n + d_n$, $d_n \in [-\epsilon, \epsilon]^m$ for all $n \in \mathbb{N}$, and y appears in the state trajectory. We construct a new state trajectory $(y_0, y_1, \ldots)$ and the sequence $(\bar{d}_0, \bar{d}_1, \ldots)$ with the transformation $x_n = Qy_n$ and $d_n = Q\bar{d}_n$. Then we have $y_{n+1} = Q^{-1}AQy_n + Q^{-1}d_n = By_n + \bar{d}_n$. Note that $\left\|\bar{d}_n\right\|_\infty = \left\|Q^{-1}d_n\right\|_\infty \leq \left\|Q^{-1}\right\|_\infty \left\|d_n\right\|_\infty \leq \gamma_1$. Since y appears in the state trajectory $(x_0, x_1, \ldots)$, $Q^{-1}y$ appears in the state trajectory $(y_0, y_1, \ldots)$ with $y_0 = Q^{-1}x_0 = Q^{-1}x$. Therefore, $Q^{-1}y \in \widetilde{\mathcal{O}}_{\gamma_1}(B, Q^{-1}x)$ which results in $y \in Q\widetilde{\mathcal{O}}_{\gamma_1}(B, Q^{-1}x)$.

To prove the left-hand side of (4.1), we invoke the right-hand side by replacing (A, B, Q, x, ϵ) with $(B, A, Q^{-1}, Q^{-1}x, \gamma_2)$. This gives $\widetilde{\mathcal{O}}_{\gamma_2}(B, Q^{-1}x) \subseteq Q^{-1}\,\widetilde{\mathcal{O}}_{\gamma_1'}(A, x)$ with $\gamma_1' = \gamma_2 \left\|Q\right\|_\infty$. Setting $\gamma_1' = \epsilon$ proves the left-hand side of (4.1).

To prove that $\widetilde{\mathcal{O}}(A, x) = Q\widetilde{\mathcal{O}}(B, Q^{-1}x)$, we take intersection of all the sides in (4.1) over $\epsilon > 0$:

$$\bigcap_{\epsilon>0} Q\,\widetilde{\mathcal{O}}_{\gamma_2}(B, Q^{-1}x) \subseteq \bigcap_{\epsilon>0} \widetilde{\mathcal{O}}_{\epsilon}(A, x) \subseteq \bigcap_{\epsilon>0} Q\,\widetilde{\mathcal{O}}_{\gamma_1}(B, Q^{-1}x).$$

Due to the linear relation between γ_1 and γ_2 with ϵ, we get

$$Q\,\widetilde{\mathcal{O}}(B, Q^{-1}x) \subseteq \widetilde{\mathcal{O}}(A, x) \subseteq Q\,\widetilde{\mathcal{O}}(B, Q^{-1}x) \quad \Rightarrow \quad \widetilde{\mathcal{O}}(A, x) = Q\,\widetilde{\mathcal{O}}(B, Q^{-1}x).$$

□

We will use Proposition 4.6 with matrix A represented using the *Jordan canonical form.*

Fix a matrix A and let J be the real Jordan form for A. Proposition 4.6 shows that $\widetilde{\mathcal{O}}(A, x)$ can be obtained from the pseudo-orbit $\widetilde{\mathcal{O}}(J, x)$.

Our decidability proof involves a case analysis on the modulus of the eigenvalues of J. We first consider the cases where J is a single block, i.e.,

$$J = \begin{bmatrix} \Lambda & I & & \\ & \Lambda & \ddots & \\ & & \ddots & I \\ & & & \Lambda \end{bmatrix} \text{ with } \Lambda = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \text{ and } I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ or } \Lambda = [r] \text{ and } I = [1], \tag{4.2}$$

with real matrix entries $a, b, r \in \mathbb{R}$.

We shall case split on the spectral radius $\rho(J)$, which is the absolute value of the unique eigenvalue of the Jordan block J. We consider three cases: $\rho(J) < 1$, $\rho(J) > 1$ and $\rho(J) = 1$.

The following lemma will be useful in relating the first and second cases. Its proof is intuitive by reversing time.

Lemma 4.7 (Reversibility Lemma)**.** *For any invertible matrix $A \in \mathbb{R}^{m\times m}$, $x \in \widetilde{\mathcal{O}}_\epsilon(A, s)$ implies $s \in \widetilde{\mathcal{O}}_\gamma(A^{-1}, x)$ with $\gamma = \epsilon \left\|A^{-1}\right\|_\infty$. Moreover,*

$$x \in \widetilde{\mathcal{O}}(A, s) \iff s \in \widetilde{\mathcal{O}}(A^{-1}, x). \tag{4.3}$$

Proof. Any $t \in \widetilde{\mathcal{O}}_\epsilon(A, s)$ is of the form $t = A^n s + \sum_{i=0}^{n-1} A^i d_{n-i-1}$ for some $n \in \mathbb{N}$ and some d_i with $\|d_i\|_\infty \le \epsilon$. This means $s = A^{-n}t + \sum_{i=0}^{n-1} A^{-i} d'_{n-i-1}$ with $d'_{n-1-i} = A^{-1}d_i$. Since $\left\|d'_{n-1-i}\right\|_\infty \le \left\|A^{-1}\right\|_\infty \epsilon$, we get $s \in \widetilde{\mathcal{O}}_\gamma(A^{-1}, t)$. To get (4.3), notice that

$$t \in \widetilde{\mathcal{O}}(A, s) \Rightarrow t \in \bigcap_{\epsilon>0} \widetilde{\mathcal{O}}_\epsilon(A, s) \Rightarrow s \in \bigcap_{\gamma>0} \widetilde{\mathcal{O}}_\gamma(A^{-1}, t) \Rightarrow s \in \widetilde{\mathcal{O}}(A^{-1}, t).$$

Applying the same argument to the matrix A^{-1} will give the other side of (4.3). □

We now consider the first case: where the eigenvalues of J are inside the unit circle.

Lemma 4.8 (Eigenvalues inside the unit circle)**.** *Let $J \in \mathbb{R}^{m\times m}$ be a Jordan block of the form (4.2) with $\rho(J) < 1$. For every $s \in \mathbb{R}^m$,*

$$\widetilde{\mathcal{O}}(J, s) = \mathcal{O}(J, s) \cup \{\mathbf{0}\} = \overline{\mathcal{O}(J, s)},$$

where $\overline{\mathcal{O}(J, s)}$ denotes the closure of the orbit.

Proof. We prove the lemma by showing there is a constant $C > 0$ satisfying

$$\overline{\mathcal{O}(J,s)} \overset{*}{=} \mathcal{O}(J,s) \cup \{\mathbf{0}\} \overset{**}{\subseteq} \widetilde{\mathcal{O}}(J,s) \overset{\dagger}{\subseteq} \bigcap_{\epsilon>0} \bigcup_{z \in \mathcal{O}(J,s)} \mathcal{B}(z, C\epsilon) \overset{\S}{\subseteq} \overline{\mathcal{O}(J,s)}, \tag{4.4}$$

where $\mathcal{B}(z,\epsilon) := \{y \in \mathbb{R}^m \mid \|z-y\|_2 \leq \epsilon\}$ is the closed ball with respect to two norm with center z and radius ϵ. It is easy to see that equality (*) holds since all the eigenvalues of J are inside the unit circle, $\lim_{n\to\infty} J^n = 0$, and 0 is the only limiting point of any state trajectory.

It is also easy to see that inclusion (**) is correct. Note that for any $\epsilon > 0$, $\mathcal{O}(J,s) \subseteq \widetilde{\mathcal{O}}_\epsilon(J,s)$ and the set $\widetilde{\mathcal{O}}_\epsilon(J,s)$ is closed by definition. Taking intersection over $\epsilon > 0$, we get $\mathcal{O}(J,s) \subseteq \widetilde{\mathcal{O}}(J,s)$ with $\widetilde{\mathcal{O}}(J,s)$ being a closed set. Therefore, $\overline{\mathcal{O}(J,s)} \subseteq \widetilde{\mathcal{O}}(J,s)$.

We now choose a value of C which allows us to prove inclusion (†). First pick γ such that $\rho(J) < \gamma < 1$. Next choose c_1 to be a constant (which is guaranteed to exist by Gelfand's formula) satisfying $\|J^n\|_2 \leq c_1\gamma^n$ for all $n \in \mathbb{N}$, and finally set $C := c_1 m/(1-\gamma)$. We show that $\widetilde{\mathcal{O}}_\epsilon(J,s) \subseteq \bigcup_{z\in\mathcal{O}(J,s)} \mathcal{B}(z, C\epsilon)$ for any $\epsilon > 0$. Take any $x \in \widetilde{\mathcal{O}}_\epsilon(J,s)$. Then there is a sequence $(d_0, d_1, \ldots)$ and $n \in \mathbb{N}$ such that $\|d_i\|_\infty \leq \epsilon$ and $x = J^n s + \sum_{i=0}^{n-1} J^i d_{n-i-1}$. Now

$$\|x - J^n s\|_2 = \left\| \sum_{i=0}^{n-1} J^i d_{n-i-1} \right\|_2 \leq \sum_{i=0}^{n-1} \left\|J^i\right\|_2 \|d_{n-i-1}\|_2 \leq \sum_{i=0}^{n-1} c_1\gamma^i m\epsilon \leq \frac{c_1 m\epsilon}{1-\gamma} = C\epsilon,$$

We then get $x \in \mathcal{B}(z, C\epsilon)$ for $z := J^n s \in \mathcal{O}(J,s)$.

The inclusion § can be proven by taking an arbitrary point $y \notin \overline{\mathcal{O}(J,s)}$ and showing that there is an $\epsilon > 0$ for which $y \notin \mathcal{B}(z, C\epsilon)$ for all $z \in \mathcal{O}(J,s)$. Note that the complement of $\overline{\mathcal{O}(J,s)}$ is an open set, which means there is a $\theta > 0$ such that $\mathcal{B}(y,\theta) \cap \overline{\mathcal{O}(J,s)} = \emptyset$. Taking ϵ such that $C\epsilon < \theta$ will give the intended result. □

Additionally, we prove the following lemma (that will be useful later) about the behaviour of pseudo-orbits when all eigenvalues are inside the unit circle.

Lemma 4.9 (Shrinking pseudo-orbits are trapped)**.** *Let $A \in \mathbb{R}^{m\times m}$ and $s \in \mathbb{R}^m$. If $\rho(A) < 1$, then for every $\delta > 0$ there exists an effectively computable $N \in \mathbb{N}$ and $\epsilon > 0$ such that after time N, all ϵ-pseudo-orbits are contained inside the ball $\mathcal{B}(\mathbf{0}, \delta)$.*

Proof. Let $(x_n)_{n\in\mathbb{N}}$ denote an ϵ-pseudo-orbit starting from s with a sequence of disturbances $(d_n)_{n\in\mathbb{N}}$. Suppose $\rho(A) < 1$ and let $\gamma \in (\rho(A), 1)$. There is a constant $c > 0$

satisfying $\|A^n\|_2 \leq c\gamma^n$ for all n. Then we get

$$\begin{aligned}\|x_n\|_2 = \left\| A^n s + \sum_{k=0}^{n-1} A^k d_{n-k-1} \right\|_2 &\leq \|A^n\|_2 \|s\|_2 + \sum_{k=0}^{n-1} \|A^k\|_2 \|d_{n-k-1}\|_2 \\ &\leq c\gamma^n \|s\|_2 + \sum_{k=0}^{n-1} m\,\epsilon\, c\gamma^k \leq c\gamma^n \|s\|_2 + \frac{m\,\epsilon\, c}{1-\gamma}.\end{aligned}$$

Taking $\epsilon = \delta(1-\gamma)/(2mc)$ and N with $\gamma^N \|s\|_2 \leq \delta/(2c)$ gives the intended result. □

We now consider the second case.

Lemma 4.10 (Eigenvalues outside the unit circle)**.** *Let $J \in \mathbb{R}^{m\times m}$ be a Jordan block of the form* (4.2) *with $\rho(J) > 1$. For every $s \in \mathbb{R}^m$, if $s \neq \mathbf{0}$, then $\widetilde{\mathcal{O}}(J, s) = \mathcal{O}(J, s)$, whereas $\widetilde{\mathcal{O}}(J, \mathbf{0}) = \mathbb{R}^m$.*

Proof. In this case, J is invertible and all eigenvalues of J^{-1} are inside the unit circle. We apply the Reversibility Lemma 4.7 and Lemma 4.8.

$$x \in \widetilde{\mathcal{O}}(J, s) \iff s \in \widetilde{\mathcal{O}}(J^{-1}, x) \iff s \in \mathcal{O}(J^{-1}, x) \cup \{\mathbf{0}\} \iff s = \mathbf{0} \text{ or } x \in \mathcal{O}(J, s).$$

Therefore, any x is in $\widetilde{\mathcal{O}}(J, s)$ if $s = 0$, and $\widetilde{\mathcal{O}}(J, s) = \mathcal{O}(J, s)$ for $s \neq \mathbf{0}$. □

Now the third case.

Lemma 4.11 (Eigenvalues on the unit circle)**.** *Let $J \in \mathbb{R}^{m\times m}$ be a Jordan block of the form* (4.2) *with $\rho(J) = 1$. For every $s \in \mathbb{R}^m$, we have $\widetilde{\mathcal{O}}(J, s) = \mathbb{R}^m$.*

Proof. The key part of the proof is to show that $\mathbf{0} \in \widetilde{\mathcal{O}}(A, s)$ for any s and for any A having the eigenvalues on the unit circle. Once we show this, we know that $s \in \widetilde{\mathcal{O}}(A^{-1}, \mathbf{0})$ is true for any s and any matrix A due to the Reversibility Lemma 4.7. Stated for the inverse of A and any x, we get $x \in \widetilde{\mathcal{O}}(A, \mathbf{0})$. Since pseudo-orbits are transitive, we have $x \in \widetilde{\mathcal{O}}(A, s)$ for any x and s, which is the intended result.

We show $\mathbf{0} \in \widetilde{\mathcal{O}}(A, s)$ equivalently by replacing A with its Jordan form J and using induction on the structure of J. The proof has two stages. The first stage is to show that $\mathbf{0} \in \widetilde{\mathcal{O}}(J, s)$ for all s when J has a single block of size 1. The second stage is to show that we can sequentially increase the multiplicity of eigenvalues and multiple blocks.

Base case: Suppose $J = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ with $a^2+b^2 = 1$ or $J = r$ with $|r| = 1$. Observe that the multiplication by J does not increase the two norm of a vector. Hence setting

$$d_n = \begin{cases} -\epsilon \cdot \frac{Jx_n}{\|Jx_n\|_2} & \text{if } ||Jx_n||_\infty > \epsilon, \\ -Jx_n & \text{otherwise,} \end{cases}$$

we obtain the ϵ-pseudo-orbit $(x_0 = s, x_1, x_2, \ldots, x_m, \mathbf{0}, \mathbf{0}, \ldots)$ from any s where $\|x_k\|_2 = \|x_{k-1}\|_2 - \epsilon$ for $k \leq m$, which gives $\mathbf{0} \in \widetilde{\mathcal{O}}(J, s)$.

Inductive case: We show that if $\mathbf{0} \in \widetilde{\mathcal{O}}(J_1, s_1)$ and $\mathbf{0} \in \widetilde{\mathcal{O}}(J_2, s_2)$ for all s_1 and s_2 of appropriate dimensions, we also have $\mathbf{0} \in \widetilde{\mathcal{O}}(J, s)$ with $J = \begin{bmatrix} J_1 & B \\ 0 & J_2 \end{bmatrix}$ for any B and any s with appropriate dimensions. Let us partition any state $x = (x^1, x^2)$ according to the dimensions of J_1 and J_2. Let $\epsilon > 0$ and $s = (s^1, s^2)$. By the assumption, there exist ϵ-perturbations $(d_0^2, d_1^2, \ldots, d_{N-1}^2)$ that bring s^2 to $\mathbf{0}$ under J_2. Let $d_n = (\mathbf{0}, d_n^2)$ for $0 \leq n < N$ be a sequence of ϵ-perturbations for the linear system with mapping J. We obtain the sequence $(x_0 = s, x_1, \ldots, x_N)$ with $x_N^2 = \mathbf{0}$: the ϵ-perturbations $d_0, \ldots, d_{N-1}$ have brought the second coordinate to $\mathbf{0}$. By the assumption, we also have $\mathbf{0} \in \widetilde{\mathcal{O}}_\epsilon(J_1, x_N^1)$, which gives ϵ-perturbations $(d_0^1, \ldots, d_M^1)$ that bring x_N^1 to $\mathbf{0}$ under J_1. Let us expand the sequence of perturbations for the linear system J with $d_{n+N} = (d_n^1, \mathbf{0})$ for $0 \leq n \leq M$. It is easy to see that $(d_0, \ldots, d_{N+M})$ bring the system from s to $\mathbf{0}$ due to the upper triangular structure of J.

$\square$

We now consider the general case where J has multiple blocks.

Definition 4.1. Let $J \in \mathbb{R}^{m \times m}$ be a real Jordan block matrix and $s \in \mathbb{R}^m$. We define

$$\Delta(J, s) := \begin{cases} \mathbb{R}^m & \text{if } \rho(J) = 1 \text{ or, } \rho(J) > 1 \text{ and } s = \mathbf{0}, \\ \{\mathbf{0}\} & \text{if } \rho(J) < 1, \\ \emptyset & \text{otherwise.} \end{cases}$$

The following lemma states that certain points in the pseudo-orbit of real Jordan blocks are ϵ-pseudo reachable exactly at any sufficiently large time step, for every $\epsilon > 0$. The lemma provides the flexibility to "synchronize" reaching parts of the state for different Jordan blocks.

Lemma 4.12 (Synchronization Lemma)**.** *Let $J \in \mathbb{R}^{m \times m}$ be a Jordan block with eigenvalue λ. For $s \in \mathbb{R}^m$, $t \in \Delta(J, s)$ if and only if for every $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that for all $N > N_\epsilon$, there exists an ϵ-pseudo-orbit $(x_i)_{i \in \mathbb{N}}$ of s under J such that $x_N = t$.*

Proof. We work case by case.

- $|\lambda| < 1$ and $\Delta(J, s) = \{\mathbf{0}\}$. By Lemma 4.8, $\mathbf{0} \in \widetilde{\mathcal{O}}(J, s)$ and hence for every $\epsilon > 0$, there exists N_ϵ such that $t = \mathbf{0}$ can be ϵ-pseudo reached at time N_ϵ. Now simply observe that once an ϵ-pseudo-orbit reaches $\mathbf{0}$, it can remain there forever by setting all future perturbations to zero. To prove the other direction, suppose $t \neq \mathbf{0}$. By Lemma 4.9, there must exist a time bound T such that for sufficiently small ϵ, all ϵ-pseudo-orbits of s after time T are contained in $\mathcal{B}(\mathbf{0}, \frac{\|t\|_2}{2})$. Hence for sufficiently small ϵ no N_ϵ with the the specified property can exist.

- $|\lambda| = 1$ and $\Delta(J, s) = \mathbb{R}^m$. In the proof of Lemma 4.11, for every $t \in \mathbb{R}^m$ and $\epsilon > 0$ we construct an ϵ-pseudo-orbit from s that visits $\mathbf{0}$ followed by t. Let N_ϵ be the number of steps required to ϵ-reach t. We can postpone visiting t to any time step $N > N_\epsilon$ by simply waiting at the point $\mathbf{0}$ for $N - N_\epsilon$ steps.

- $|\lambda| > 1$, $s = \mathbf{0}$ and $\Delta(J, s) = \mathbb{R}^m$. Similarly to the case above, in Lemma 4.10 for each ϵ we construct an ϵ-pseudo-orbit that visits t at time N_ϵ, and reaching t can be delayed arbitrarily by spending a necessary number of steps at $\mathbf{0}$ at the beginning.

- $|\lambda| > 1$, $s \neq \mathbf{0}$ and $\Delta(J, s) = \emptyset$. Let $t \in \mathbb{R}^m$. In this case, observe that there must exist a time bound T such that for sufficiently small ϵ, all ϵ-pseudo-orbits of s after time T are contained outside $\mathcal{B}(\mathbf{0}, 2\|t\|_2)$. Hence for sufficiently small ϵ no N_ϵ with the the specified property can exist. □

There are two modes of pseudo reachability: via orbit, or at larger and larger time steps for smaller ϵ.

Lemma 4.13. *Let $A \in \mathbb{R}^{m\times m}$ and $s, t \in \mathbb{R}^m$. If there exists N such that for every ϵ, t is ϵ-pseudo-reachable from s within the first N steps, then $t \in \mathcal{O}(A, s)$.*

Proof. Suppose such N exists. By continuity of the map $x \mapsto Ax$, for every $\delta > 0$ there exists $\epsilon > 0$ such that for every $\epsilon' < \epsilon$ and ϵ'-pseudo-orbit $(x_i)_{i\in\mathbb{N}}$, $\|x_i - A^i s\|_2 < \delta$ for $0 \le i < N$. Hence the intersection of the first N elements of all ϵ-pseudo-orbits is exactly $\{s, As, \dots, A^{N-1}s\}$. □

Lemma 4.14. *For $J = \mathrm{diag}(J_1, \dots, J_l)$ in real Jordan normal form and $s \in \mathbb{R}^m$,*

$$\widetilde{\mathcal{O}}(J, s) = \mathcal{O}(J, s) \cup \Pi_{i=1}^{l} \Delta(J_i, s_i).$$

Proof. Suppose $t = (t_1, \dots, t_l) \in \Pi_{i=1}^{l} \Delta(J_i, s_i)$. That is, for every ϵ and $1 \le i \le l$ there exists an ϵ-pseudo-orbit $(x^i_j)_{j\in\mathbb{N}}$ of s_i under J_i that reaches t_i. By Lemma 4.12, for every ϵ there exist ϵ-pseudo-orbits $(y^i_j)_{j\in\mathbb{N}}$ of $s_1, \cdots, s_l$ that reach $t_1, \dots, t_l$, respectively, at the same time N. That is, $y^i_N = t_i$ for $1 \le i \le m$. Hence $(y^1_i, \dots, y^z_i)_{i\in\mathbb{N}}$ is an ϵ-pseudo-orbit of s under J that reaches t.

Now suppose $t \in \widetilde{\mathcal{O}}(J,s) \setminus O(J,s)$. We prove, by a case analysis on J_i, that $t_i \in \Delta(J_i, s_i)$ for $1 \leq i \leq l$. The main idea is that if t is pseudo-reachable but not reachable, then in order to reach it via an ϵ-pseudo-orbit one will need longer and longer time horizons as $\epsilon \to 0$ (Lemma 4.13).

1. $\rho(J_i) < 1$. Since t is not in the orbit, we can find a sequence $N_1 < N_2 < \cdots$ of time steps and $\epsilon_1 > \epsilon_2 > \cdots$ of perturbations such that t is ϵ_j-reachable from s earliest at time N_j. In particular, t_i is ϵ_j reachable from s_i at time N_j for every j. But by Lemma 4.9 this means that $|t_i| < \delta$ for every $\delta > 0$. Hence $t_i = \mathbf{0} \in \Delta(J_i, s_i)$.

2. $\rho(J_i) = 1$. Since in this case $\Delta(J_i, s_i) = \mathbb{R}^{\kappa(i)}$, trivially $t_i \in \Delta(J_i, s_i)$.

3. $\rho(J_i) > 1$ and $s_i = \mathbf{0}$. Since in this case too $\Delta(J_i, s_i) = \mathbb{R}^{\kappa(i)}$, trivially $t_i \in \Delta(J_i, s_i)$.

4. $\rho(J_i) > 1$ and $s_i \neq \mathbf{0}$. This case cannot arise, as similarly to Case 1, one can argue that if pseudo-reaching t_i requires larger and larger time steps as $\epsilon \to 0$, then $|t_i| > \delta$ for every δ. But in this case no such t_i can exist. $\square$

We now have everything we need to solve the pseudo-orbit problem.

Theorem 4.15. *Given $A \in \mathbb{Q}^{m \times m}$, and $s, t \in \mathbb{Q}^m$, we can decide in polynomial time whether $t \in \widetilde{\mathcal{O}}(A, s)$.*

Proof. Given $A \in \mathbb{Q}^{m \times m}$, and $s, t \in \mathbb{Q}^m$, we compute (in polynomial time) matrices $Q, J, Q^{-1} \in (\mathbb{R} \cap \overline{\mathbb{Q}})^{m \times m}$ such that $A = QJQ^{-1}$ and J is in real Jordan normal form (Section 2.3.1). Then, we compute $t' = Q^{-1}t$ and $s' = Q^{-1}s$, and by Proposition 4.6 we have that $t \in \widetilde{\mathcal{O}}(A, s)$ if and only if $t' \in \widetilde{\mathcal{O}}(J, s')$. It remains to decide whether $t' \in \widetilde{\mathcal{O}}(J, s')$. For this we use the characterization described in Lemma 4.14. To decide whether $t' \in \mathcal{O}(J, s')$, observe that $Q^{-1}t \in \mathcal{O}(J, Q^{-1}s) \iff t \in \mathcal{O}(A, s)$, and whether $t \in \mathcal{O}(A, s)$ is an instance of the Orbit Problem and can be decided in polynomial time. Finally, it remains to check whether $t_i \in \Delta(J_i, s_i)$ for each block J_i, which can be done easily given the simplicity of $\Delta(J_i, s_i)$. $\square$

4.4 Pseudo-reachability is decidable for bounded semialgebraic sets

We now apply our characterization of pseudo-orbits (Lemma 4.14) to deciding pseudo-reachability for arbitrary bounded semialgebraic sets.

Definition 4.2 (Pseudo-reachability)**.** A set S is *pseudo-reachable* from s under A if for every $\epsilon > 0$, there exists a point $x_\epsilon \in S$ that is ϵ-pseudo-reachable from s under A.

Note that the quantifier order is important: a fixed point that is ϵ-pseudo-reachable for all $\epsilon > 0$ must be in the pseudo-orbit. For (closed) bounded sets, we will show that these two notions coincide, but (as we will see in the next section), it is possible to have closed unbounded sets which are pseudo-reachable despite not intersecting the pseudo-orbit.

Recall that a *semialgebraic set* is the solution set of a boolean combination of integer polynomial equalities and inequalities. We show that we can decide if a bounded semialgebraic set is pseudo-reachable, by reducing the problem to checking if the closure of the set intersects the pseudo-orbit. The following proposition shows that we can decide if a closed bounded semialgebraic set intersects a pseudo-orbit.

Proposition 4.16. *Given $A \in \mathbb{Q}^{m\times m}$, $x_0 \in \mathbb{Q}^m$, and a closed bounded semialgebraic set C, it is decidable if C intersects $\widetilde{\mathcal{O}}(A, x_0)$.*

Proof. Convert A to Jordan form J. Recall that for $J = \operatorname{diag}(J_1, \ldots, J_l)$ in real Jordan normal form and $s \in \mathbb{R}^m$,

$$\widetilde{\mathcal{O}}(J, s) = \mathcal{O}(J, s) \cup \Pi_{i=1}^{l} \Delta(J_i, s_i),$$

where

$$\Delta(J, s) := \begin{cases} \mathbb{R}^\kappa & \text{if } \rho(J) = 1 \text{ or, } \rho(J) > 1 \text{ and } s = \mathbf{0}, \\ \{\mathbf{0}\} & \text{if } \rho(J) < 1, \\ \emptyset & \text{otherwise.} \end{cases}$$

We proceed casewise based on the value of $\Delta := \Pi_{i=1}^{l} \Delta(J_i, s_i)$.

Case 1: $\Delta = \emptyset$. In this case we know that the orbit must blow up due to the presence of a Jordan block with eigenvalue $\lambda_i > 1$ and a non-zero corresponding component s_i. We can find R such that $C \subseteq \mathcal{B}(\mathbf{0}, R)$. Set $N = \lceil (\log R - \log |s_i|)/\log \lambda_i \rceil$. Clearly for $n > N$, $A^n x_0 \notin C$, so we only need to check whether the first N points of the orbit intersect C.

Case 2: $\Delta \neq \emptyset$. In this case, without loss of generality we can write $\Delta = \mathbf{0}^k \times \mathbb{R}^{m-k}$ for some k, $0 \leq k \leq m$. Checking whether this Δ intersects C is trivial. If it does not intersect, it is because C does not touch zero in all the zero dimensions. Let C' be the projection of C onto the first k dimensions. Since C' is a compact semialgebraic set that does not contain zero, using Lemma 2.15 we can find δ such that $C' \subseteq \mathbb{R}^k \setminus \mathcal{B}(\mathbf{0}^k, \delta)$. By Lemma 4.9, we can effectively compute $N \in \mathbb{N}$ and $\epsilon > 0$ such that after time N, all ϵ-pseudo-orbits are contained inside $\mathcal{B}(\mathbf{0}^k, \delta) \times \mathbb{R}^{m-k}$. Thus again we only need to check the first N points of the orbit. □

Theorem 4.17. *Given $A \in \mathbb{Q}^{m \times m}$, $x_0 \in \mathbb{Q}^m$, and a bounded semialgebraic set S, it is decidable if S is pseudo-reachable from x_0 under A.*

Proof. We show that S is pseudo-reachable from x_0 under A if and only if there exists $x \in \overline{S}$ that is pseudo-reachable from x_0 under A, allowing us to restrict our attention to compact sets and the existence of a pseudo-reachable point in a set as opposed to pseudo-reachability of the set as a whole. Deciding pseudo-reachability then reduces to checking whether $\overline{S} \cap \widetilde{\mathcal{O}}(A, x_0) = \emptyset$, which can be computed using Proposition 4.16.

Suppose S is pseudo-reachable. Let $(\epsilon_i)_{i \in \mathbb{N}}$ be a sequence of positive numbers with $\lim_{i \to \infty} \epsilon_i = 0$, and $(x_i)_{i \in \mathbb{N}}$ be a sequence of elements of S such that x_i is ϵ_i-pseudo-reachable for all $i \geq 0$. By the Bolzano–Weierstrass theorem, boundedness of S implies that $(x_i)_{i \in \mathbb{N}}$ must have a limit point x in $\overline{S}$. To argue that x is pseudo-reachable, let $\epsilon > 0$. Since x is the limit point of $(x_i)_{i \in \mathbb{N}}$, there must exist an $\frac{\epsilon}{2}$-pseudo-orbit $(y_i)_{i \in \mathbb{N}}$ containing a point y_N such that $\|x - y_N\|_\infty < \frac{\epsilon}{2}$. Therefore, x is ϵ-pseudo-reachable from s via the sequence $s, y_1, \ldots, y_{N-1}, x$.

Now suppose $x \in \overline{S}$ is pseudo-reachable. To argue that S is pseudo-reachable, let $\epsilon > 0$. Since $x \in \overline{S}$, there must exist a point $x' \in S$ such that $\|x' - x\|_\infty < \frac{\epsilon}{2}$. Since x is $\frac{\epsilon}{2}$-pseudo-reachable, x' must be ϵ-pseudo-reachable. □

4.5 Hyperplane pseudo-reachability is decidable

In this section, we study the hyperplane pseudo-reachability problem, a generalization of the Skolem problem to pseudo-orbits.

Problem 4.18 (Hyperplane pseudo-reachability problem)**.** *Given $A \in \mathbb{Q}^{m \times m}$, $s \in \mathbb{Q}^m$, and a hyperplane $c^T \cdot x = v$ for $c \in \mathbb{Q}^m, v \in \mathbb{Q}$, decide whether $\widetilde{\mathcal{O}}_\epsilon(A, s)$ intersects the hyperplane for all $\epsilon > 0$.*

First we consider the case where we are given:

- a hyperplane $H = \{x \in \mathbb{R}^m : c^\top x = v\}$ with $(c, v) \in (\mathbb{R} \cap \overline{\mathbb{Q}})^m \times (\mathbb{R} \cap \overline{\mathbb{Q}})$,
- $J = \operatorname{diag}(J_1, \ldots, J_z) \in (\mathbb{R} \cap \overline{\mathbb{Q}})^{m \times m}$ in real Jordan normal form, and
- a starting point $x_0 \in (\mathbb{R} \cap \overline{\mathbb{Q}})^m$.

We show how to decide if for every $\epsilon > 0$ there exists an ϵ-pseudo-orbit $(x_i)_{i\in\mathbb{N}}$ of x_0 under J that *hits* the hyperplane H, i.e. $c^\top x_N - v = 0$ for some $N \in \mathbb{N}$.

A block J_i is *relevant* with respect to hyperplane $H = \{x : c^\top x = v\}$ if the coefficients of c at the coordinates corresponding to J_i are not all 0. Intuitively, dimensions corresponding to blocks that are not relevant can simply be omitted from the analysis as they do not play a role in determining whether a point is in H or not. *Relevant eigenvalues* of J are the eigenvalues of relevant blocks. The *relevant spectral radius*, written $\rho_H(J)$, is the largest modulus of all relevant eigenvalues. Our proof uses a case analysis on $\rho_H(J)$. We shall see that the proof is simple when $\rho_H(J) \le 1$ but requires more technical ideas when $\rho_H(J) > 1$.

Lemma 4.19 (Case $\rho_H(J) \le 1$)**.** *Fix a matrix J in real Jordan normal form, a starting state x_0, and a hyperplane $H = \{x : c^\top x = v\}$.*

1. *If $\rho_H(J) = 1$, then H is pseudo-reachable.*
2. *If $\rho_H(J) < 1$ and $\mathbf{0} \in H$ then H is pseudo-reachable. If $\rho_H(J) < 1$ and $\mathbf{0} \notin H$, there exists an effectively computable time bound N such that H is pseudo-reachable if and only if there exists $0 \le i \le N$ such that $J^i x_0 \in H$ (that is, H is reachable from x_0 under J after at most N steps).*

Proof. First suppose $\rho_H(J) = 1$. We write $J = diag(J_h, J_r)$, where $\rho_H(J_h) = 1$ and $\rho_H(J_r) < 1$ (observe that wlog we can assume the blocks of J have non-decreasing spectral radius when listed from top to bottom) and correspondingly set $s = (s_h, s_r)$, $c = (c_h, c_r)$. Note that $c_h \ne 0$ by the relevance of at least one eigenvalue of modulus 1.

By Lemma 4.8 we know $\mathbf{0} \in \widetilde{\mathcal{O}}(J_r, s_r)$. By Lemma 4.11, we can select y such that $c_h^\top y - v = 0$ and $y \in \widetilde{\mathcal{O}}(J_h, s_h)$. Therefore, invoking Lemma 4.12, for every $\epsilon > 0$ we can find $N \in \mathbb{N}$ and construct ϵ-pseudo-orbits $(x_n^h)_{n\in\mathbb{N}}$ and $(x_n^r)_{n\in\mathbb{N}}$ such that $x_N^h = y$ and $x_N^r = \mathbf{0}$, which implies that for the ϵ-pseudo-orbit $x_n = (x_n^h, x_n^r)$, $c^\top x_N - v = c_h^\top y + c_r^\top \mathbf{0} - v = 0$ as desired.

Now suppose $\rho_H(J) < 1$.

Case 1: $v = 0$**.** Since $\mathbf{0} \in H$ and the origin is pseudo-reachable from x_0 (Lemma 4.8), H is pseudo-reachable.

Case 2: $v \ne 0$**.** Using Lemma 4.9, and setting $\delta = |v|/(2\,\|c\|_2)$, we can find $\epsilon > 0$ and horizon $N \in \mathbb{N}$ after which every ϵ-pseudo-orbit is trapped in $\mathcal{B}(0, \delta)$. Thus, the hyperplane cannot be pseudo-reached after time N, as the hyperplane does not intersect with $\mathcal{B}(0, \delta)$. It remains to check if the hyperplane is pseudo-reachable at any of the first N time-steps. In fact, for a bounded time interval, a hyperplane is pseudo-reachable iff it is reachable. This is because the effect of finitely many disturbance terms $(d_0, \ldots, d_{N-1})$ can be made arbitrarily small for small enough ϵ. Therefore, decidability in this case only requires checking if the bounded orbit $(y_n)_{0 \le n \le N}$ hits the hyperplane before the time horizon N, that is, if there exists a time-step $0 \le n \le N$ such that $c^\top y_n - v = 0$, which is clearly decidable. $\square$

We now consider the case $\rho_H(J) > 1$. The main ideas of our proof are as follows:

1. A point x_n in the ϵ-pseudo orbit belongs to the hyperplane (c, v) if $c^\top x_n - v = 0$. In particular, $c^\top x_n - v$ can be written as a sum over exponential polynomials in eigenvalues of different sizes.

2. We factor out the scaling factor corresponding to the top eigenvalues, leaving a sum over normalized eigenvalues, together with a sum over disturbances (of order ϵ) and additional terms which go to zero with large n.

3. We relate hyperplane pseudo-reachability to the limit inferior of the sum over normalized eigenvalues. If the limit is zero, we show the hyperplane is pseudo-reachable. If the limit is positive, we show there is an effective bound N such that if the hyperplane is pseudo-reachable, it is reachable within N steps.

4. We apply results from Diophantine approximation and the theory of reals to compute the limit inferior of the sum over normalized eigenvalues.

Fix $J = \mathrm{diag}(J_1, \ldots, J_l) \in (\mathbb{R} \cap \overline{\mathbb{Q}})^{m \times m}$, a starting point $x_0 \in (\mathbb{R} \cap \overline{\mathbb{Q}})^m$, and a hyperplane $H = \{x \in \mathbb{R}^m \mid c^\top x = v\}$ with $c \in (\mathbb{R} \cap \overline{\mathbb{Q}})^m, v \in \mathbb{R} \cap \overline{\mathbb{Q}}$. We assume without loss of generality that all blocks are relevant.

Step 1: Analysing $c^\top x_n - v$. Let $L = \rho_H(J) > 1$ be the largest modulus of a relevant eigenvalue of J and suppose the blocks are arranged in non-increasing order of the modulus of eigenvalues. In particular, let $t \le l$ be such that the first t blocks (t for "top") have $\rho(J_1) = \cdots = \rho(J_t) = L > 1$. We call the eigenvalues of these blocks the *top eigenvalues*. The remaining blocks satisfy $L > \rho(J_{t+1}) \ge \cdots \ge \rho(J_l)$.

Let $(d_i)_{i \in \mathbb{N}}$ be a sequence of perturbations and $(x_i)_{i \in \mathbb{N}}$ the resulting pseudo-orbit. We have that for all time steps n,

$$c^\top x_n - v = c^\top \left(J^n x_0 + \sum_{k=0}^{n-1} J^k d_{n-k-1} \right) - v = \sum_{i=1}^{l} \left(c^i J_i^n x_0^i + c^i \sum_{k=0}^{n-1} J_i^k d_{n-k-1}^i \right) - v,$$

where for all $1 \leq i \leq l$, c^i, x_n^i d_n^i are projections of $c^\top$, x_n and d_n, respectively, onto the coordinates governed by J_i. Observe that c^i is a row vector for every i.

Step 2: Normalized sum. We define a normalized version of this sum by factoring out L^n (the size of the top eigenvalues) and n^D, where we define D in such a way that we normalize polynomials in n that appear in the sum. Observe that for $1 \leq i \leq t$ (the top eigenvalues),

$$c^i J_i^n = \begin{cases} \begin{bmatrix} p_1^i(n)\lambda^n + \overline{p_1^i(n)\lambda^n} & \cdots & p_{2\kappa(i)}^i(n)\lambda^n + \overline{p_{2\kappa(i)}^i(n)\lambda^n} \end{bmatrix} & \text{if } J_i \text{ has eigenvalues } \lambda, \overline{\lambda} \\ \begin{bmatrix} p_1^i(n)\rho^n & \cdots & p_{\kappa(i)}^i(n)\rho^n \end{bmatrix} & \text{if } J_i \text{ has a single eigenvalue } \rho \end{cases}$$

for polynomials $p_1^i, \ldots, p_{\kappa(i)}^i$ (with algebraic coefficients) where $\kappa(i)$ is the multiplicity of the block J_i.

We define D to be the largest number such that the monomial n^D appears with a non-zero coefficient in at least one of $c^i J_i^n$ for $1 \leq i \leq t$. (Note that if all entries of c are non-zero $D+1$ is equal to the largest multiplicity of a top eigenvalue block of J, as can be seen from the description of powers of a Jordan block in Section 2.3.)

We can now define

$$f(n) := \frac{c^\top \cdot x_n - v}{L^n n^D} = \sum_{i=1}^{l} \left(c^i \frac{J_i^n}{L^n n^D} x_0^i + c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^D} d_{n-k-1}^i \right) - \frac{v}{L^n n^D}$$

For notational convenience we define vector-valued functions $g^i(n) := c^i \frac{J_i^n}{L^n n^D}$ for $1 \leq i \leq l$. The following technical lemma summarizes the relevant properties of these scaled terms.

Lemma 4.20 (Normalization Lemma)**.**

1. *For $1 \leq i \leq t$ (top eigenvalues), $\|g^i(n)\|_\infty = O(1)$ (with respect to n).*
2. *For $t+1 \leq i \leq l$ (non-top eigenvalues), $\lim_{n\to\infty} \|g^i(n)\|_\infty = 0$.*
3. *There exists $1 \leq j \leq t$ and effectively computable $N \in \mathbb{N}$ and $C > 0$ such that $n > N \implies \|g^j(n)\|_\infty > C$.*

Proof. We address each point individually.

1: For $1 \leq i \leq t$ let J_i have eigenvalues λ and $\overline{\lambda}$ (the case where J_i has a single real eigenvalue is similar but simpler) and observe that

$$g^i(n) = \begin{bmatrix} \frac{p_1^i(n)}{n^D}\left(\frac{\lambda}{L}\right)^n + \overline{\frac{p_1^i(n)}{n^D}}\left(\frac{\overline{\lambda}}{L}\right)^n & \cdots & \frac{p_{2\kappa(i)}^i(n)}{n^D}\left(\frac{\lambda}{L}\right)^n + \overline{\frac{p_{2\kappa(i)}^i(n)}{n^D}}\left(\frac{\overline{\lambda}}{L}\right)^n \end{bmatrix}.$$

By the definition of top eigenvalues, $|\lambda| = L$ and thus $\frac{\lambda}{L}$ and $\frac{\overline{\lambda}}{L}$ have modulus 1. By construction of n^D, the polynomials $p_1^i(n), \ldots, p_{2\kappa(i)}^i$ all have degree at most D and hence the terms $\frac{p_1^i(n)}{n^D}, \ldots, \frac{p_{2\kappa(i)}^i(n)}{n^D}$ are bounded from above by a constant.

2: For $t+1 \leq i \leq l$ let J_i have eigenvalues λ and $\overline{\lambda}$ and observe that

$$g^i(n) = \begin{bmatrix} \frac{p_1^i(n)}{n^D}\left(\frac{\lambda}{L}\right)^n + \frac{\overline{p_1^i(n)}}{n^D}\left(\frac{\overline{\lambda}}{L}\right)^n & \cdots & \frac{p_{\kappa(i)}^i(n)}{n^D}\left(\frac{\lambda}{L}\right)^n + \frac{\overline{p_{\kappa(i)}^i(n)}}{n^D}\left(\frac{\overline{\lambda}}{L}\right)^n \end{bmatrix}.$$

By construction $|\lambda| < L$ and thus $\gamma := \frac{\lambda}{L}$ and $\overline{\gamma}$ have moduli $|\gamma|, |\overline{\gamma}| < 1$. The polynomials $p_1^i(n), \ldots, p_{2\kappa(i)}^i(n)$ may not be asymptotically bounded by n^D (since n^D was constructed only considering top eigenvalues). However, it is clear that the exponentially vanishing $\left(\frac{\lambda}{L}\right)^n$ and $\left(\frac{\overline{\lambda}}{L}\right)^n$ will dominate the polynomials and all entries of $g^i(n)$ will thus vanish.

3: Observe that by the construction of n^D, there must exist a top eigenvalue block J_j $(1 \leq j \leq t)$ for which at least one polynomial in $c^j J_j^n$ has degree D. Let $r > D$ be the multiplicity of the block J_j, which has the form of a real Jordan matrix with a single block (Eq. (4.2)) with sub-blocks Λ. One can write

$$c^j J_j^n = \begin{bmatrix} c_r^j & c_{r-1}^j & \cdots & c_0^j \end{bmatrix} \begin{pmatrix} \Lambda^n & n\Lambda^{n-1} & \binom{n}{2}\Lambda^{n-1} & \cdots & \binom{n}{r-1}\Lambda^{n-r+1} \\ 0 & \Lambda^n & n\Lambda^{n-1} & \cdots & \binom{n}{r-2}\Lambda^{n-r+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & n\Lambda^{n-1} \\ 0 & 0 & 0 & \cdots & \Lambda^n \end{pmatrix}, \tag{4.5}$$

where c_k^j for $1 \leq k \leq r$ corresponds to a row vector of size two or one, respectively, when Λ is a 2×2 or 1×1 matrix. Analyzing this product, we see that $c_r^j, \ldots, c_{D+1}^j = \mathbf{0}$, $c_D^j \neq \mathbf{0}$ and the single entry of $c^j J_j^n$ whose polynomial component has degree D is exactly $c_D^j \binom{n}{D}\Lambda^{n-D}$.

We define $\hat{\Lambda} := \Lambda/L$. Note that $\left\|\hat{\Lambda}\right\|_2 = 1$. Now observe that for this block J_j, we have

$$g^j(n) = \frac{1}{L^n n^D} c^j J_i^n = c_D^j \frac{1}{D!}\hat{\Lambda}^n + \frac{1}{n}(O(1)).$$

Therefore, there exists sufficiently large N such that for all $n \in \mathbb{N}$,

$$n > N \implies \left\| \frac{1}{L^n n^D} c^j J_j^n \right\|_\infty > \frac{1}{2}\left\| c_D^j \frac{1}{D!}\hat{\Lambda}^n \right\|_\infty > \frac{\|c_D^j\|_2}{4D!}.$$

Thus we have shown Point 3 with $C = \frac{\|c_D^j\|_2}{4D!}$. $\square$

Step 3: Conditions for reachability and non-reachability. Now we are ready to attack our original problem. Going back, H is ϵ-pseudo-reachable if and only if

$f(n) = 0$ for some disturbance sequence $(d_i)_{i\in\mathbb{N}}$ with $d_i \in [-\epsilon, \epsilon]^m$ for all i. We analyze how $f(n)$ can be brought to 0 in this way.

Lemma 4.21. *Let*

$$G = \liminf_{n\to\infty} \left| \sum_{i=1}^{t} g^i(n) x_0^i \right|. \tag{4.6}$$

If $G = 0$, then H is pseudo-reachable. If $G > 0$, there exists a computable time bound N such that H is pseudo-reachable if and only if it is reachable (in the standard sense) within the first N steps.

Proof. Suppose $G = 0$. Take an arbitrary $\epsilon > 0$. We argue that H is ϵ-pseudo-reachable. Recall that

$$\begin{aligned} f(n) &= \sum_{i=1}^{l} \left(c^i \frac{J_i^n}{L^n n^D} x_0^i + c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^D} d_{n-k-1}^i \right) - \frac{v}{L^n n^D} \\ &= \sum_{i=1}^{l} \left(g^i(n) x_0^i + c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^D} d_{n-k-1}^i \right) - \frac{v}{L^n n^D}. \end{aligned}$$

Let I be such that $\left\| g^I(n) \right\|_\infty > C$, for $C > 0$ and sufficiently large n (Point 3 of the Normalization Lemma). We construct a pseudo-orbit with all perturbations set to zero except d_0^I and obtain

$$f(n) = c^I \frac{J_I^{n-1}}{L^n n^D} d_0^I + \sum_{i=1}^{t} g^i(n) x_0^i + \sum_{i=t+1}^{l} c^i \frac{J_i^n}{L^n n^D} x_0^i - \frac{v}{L^n n^D}.$$

Intuitively, we will use the term $c^I \frac{J_I^{n-1}}{L^n n^D} d_0^I$ to cancel out the remaining summands above, but we have to argue that this can be done using a disturbance of size at most ϵ. Moreover, observe that $c^I \frac{J_I^{n-1}}{L^n n^D}$ is very close to $g^I(n)$. Formally, we first find N large enough such that

- $\left\| g^I(N) \right\|_\infty > C$,
- $\left\| \sum_{i=t+1}^{l} c^i \frac{J_i^N}{L^N N^D} x_0^i - \frac{v}{L^N N^D} \right\|_\infty < \frac{C^2}{\|J_i\|_\infty} \frac{\epsilon}{2}$ (possible because for $t+1 \le i \le l$, $\rho(J_i) < 1$ and $L > 1$), and
- $\left\| \sum_{i=1}^{t} g^i(N) x_0^i \right\|_\infty < \frac{C^2}{\|J_i\|_\infty} \frac{\epsilon}{2}$ (possible because $\liminf_{n\to\infty} \left\| \sum_{i=1}^{t} g^i(n) x_n^i \right\| = 0$).

Finally, we determine the value of d_0^I. Without loss of generality, assume that $g^I(N)$ is of the form $\begin{bmatrix} C' & \cdots \end{bmatrix}$ where $|C'| > C$, that is the first entry of $g^I(N)$ is large. We then observe that $c^I \frac{J_I^{N-1}}{L^N N^D} d_0^I = g^I(N) J_I^{-1} d_0^I$ and set

$$d_0^I = J_I \cdot \begin{bmatrix} -\frac{1}{C'} \left(\sum_{i=1}^{t} g^i(N) x_0^i + \sum_{i=t+1}^{l} c^i \frac{J_i^N}{L^N N^D} x_0^i - \frac{v}{L^N N^D} \right) & 0 & 0 & \cdots & 0 \end{bmatrix}^\top$$

to obtain

$$c^I \frac{J_I^{N-1}}{L^N N^D} d_0^I = -\left(\sum_{i=1}^{t} g^i(N) x_0^i + \sum_{i=t+1}^{l} c^i \frac{J_i^N}{L^N N^D} x_0^i - \frac{v}{L^N N^D} \right)$$

and hence $f(N) = 0$.

Now suppose $G > 0$. Recall

$$\begin{aligned} f(n) &= \sum_{i=1}^{l} \left(g^i(n) x_0^i + c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^D} d_{n-k-1}^i \right) - \frac{v}{L^n n^D} \\ &= \sum_{i=1}^{t} g^i(n) x_0^i + \sum_{i=t+1}^{l} c^i \frac{J_i^n}{L^n n^D} x_0^i + \sum_{i=1}^{l} c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^D} d_{n-k-1}^i - \frac{v}{L^n n^D}. \end{aligned}$$

In this case we shall construct a time bound N after which for all sufficiently small value of ϵ, the term $\sum_{i=1}^{t} g^i(n) x_0^i$ will dominate the other summands. Let $2\Delta > 0$ be a lower bound on $\liminf_{n\to\infty} |\sum_{i=1}^{t} g^i(n) x_0^i| > 0$. We shall see how to obtain such a bound effectively later (Lemma 4.22). We compute N with the following properties.

- For all $n > N$, $\left|\sum_{i=1}^{t} g^i(n) x_0^i\right| > \Delta$. Possible because $\liminf_{n\to\infty} \left|\sum_{i=1}^{t} g^i(n) x_n^i\right| > 2\Delta$.
- For all $n > N$, $\left|\sum_{i=t+1}^{l} c^i \frac{J_i^n}{L^n n^d} x_0^i\right|, \left|\frac{v}{L^n n^d}\right| \ll \Delta$. The former is possible because for $t+1 \le i \le l$, $\rho(J_i) < L$.
- For sufficiently small ϵ, for all $n > N$, $\left|c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^d} d_{n-k-1}^i\right| \ll \Delta$ for $1 \le i \le l$. To see that this is always possible, observe that

$$\left| c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^d} d_{n-k-1}^i \right| \le \sum_{k=0}^{n-1} \left\| c^i \frac{J_i^k}{L^n n^d} \right\|_\infty M\epsilon$$

(where fixed M bounds the matrix dimension), and

$$\lim_{n\to\infty} \sum_{k=0}^{n} \left\| c^i \frac{J_i^k}{L^n n^d} \right\|_\infty \le \lim_{n\to\infty} \sum_{k=0}^{n} \left\| c^i \frac{1}{L^{n-k}} \frac{J_i^k}{L^k k^d} \right\|_\infty = \lim_{n\to\infty} \sum_{k=0}^{n} \left\| \frac{1}{L^{n-k}} g^i(k) \right\|_\infty.$$

Recalling Point 1 of the Normalization Lemma, $\|g^i(n)\|_\infty = O(1)$ and hence

$$\lim_{n\to\infty} \sum_{k=0}^{n} \left\| \frac{1}{L^{n-k}} g^i(k) \right\|_\infty = O(1),$$

by bounding the sum $\sum_{k=0}^{n} \left\| \frac{1}{L^{n-k}} g^i(k) \right\|_\infty$ from above by a geometric sequence. Therefore, $\sum_{k=0}^{n-1} \left\| c^i \frac{J_i^k}{L^n n^d} \right\|_\infty M\epsilon$ can be made $\ll \Delta$ by choosing ϵ to be sufficiently small.

Once we have chosen N, by the properties above we will have that for all $n > N$, for sufficiently small ϵ,

$$|f(n)| \geq \left|\sum_{i=1}^{t} g^i(n)x_0^i\right| - \left|\sum_{i=t+1}^{l} c^i \frac{J_i^n}{L^n n^d} x_0^i + \sum_{i=1}^{l} c^i \sum_{k=0}^{n-1} \frac{J_i^k}{L^n n^d} d_{n-k-1}^i - \frac{v}{L^n n^d}\right| > 0.$$

Therefore, H is pseudo-reachable if and only if for every $\epsilon > 0$, H is ϵ-pseudo-reachable within the first N steps. By Lemma 4.13, this is the case if and only if H is reachable within the first N steps. $\square$

Step 4: Analyzing $\liminf_{n\to\infty} |\sum_{i=1}^{t} g^i(n)x_0^i|$. Consider a single term $g^i(n)x_0^i$. Writing $x_0^i = \begin{bmatrix} X_0 & X_1 & \cdots & X_z \end{bmatrix}^\top$, where $X_1, \ldots, X_z \in \mathbb{R}$, we have

$$g^i(n)x_0^i = \sum_{r=1}^{z} \left(\frac{p_r^i(n)}{n^D} \left(\frac{\lambda}{L}\right)^n + \overline{\frac{p_r^i(n)}{n^D}} \left(\frac{\overline{\lambda}}{L}\right)^n \right) X_r.$$

Let $\gamma_i = \frac{\lambda}{L}$. Note that $|\gamma_i| = 1$. By the construction of n^D, none of the polynomials have a term of degree higher than D. Therefore, we can absorb the constants X_r and the monomial n^D into the polynomials, sum the terms up, and write them as polynomials in $\frac{1}{n}$. That is,

$$g^i(n)x_0^i = q^i(1/n)\gamma_i^n + \overline{q^i(1/n)}\overline{\gamma_i}^n$$

for suitable polynomials q^i with algebraic coefficients. Thus

$$\liminf_{n\to\infty} \left|\sum_{i=1}^{t} g^i(n)x_n^i\right| = \liminf_{n\to\infty} \left|\sum_{i=1}^{t} q^i(1/n)\gamma_i^n + \overline{q^i(1/n)}\overline{\gamma_i}^n\right|$$

We defer the proof of the following lemma, which requires tools from Diophantine analysis and the theory of reals, to the next section.

Lemma 4.22. *Let $\gamma_1, \ldots, \gamma_t$ be algebraic numbers with modulus 1. Let $q^1, \ldots, q^t$ be polynomials with algebraic coefficients. The quantity*

$$\liminf_{n\to\infty} \left|\sum_{i=1}^{t} q^i(1/n)\gamma_i^n + \overline{q^i(1/n)}\overline{\gamma_i}^n\right|$$

can be effectively computed. If it is greater than zero, there is an effectively computable N satisfying the requirement of Lemma 4.21.

We are now ready to aggregate our case analysis into solving the pseudo-Skolem problem.

Theorem 4.23 (Hyperplane pseudo-reachability (pseudo-Skolem) is decidable)**.** *Given $A \in \mathbb{Q}^{m\times m}$, $x_0 \in \mathbb{Q}^m$ and $H = \{x : c^\top \cdot x = v\}$ (where $c \in \mathbb{Q}^m$, $v \in \mathbb{Q}$), it is decidable whether $\widetilde{\mathcal{O}}_\epsilon(A, x_0)$ intersects H for all $\epsilon > 0$.*

Proof. Given $A \in \mathbb{Q}^{m\times m}$, $x_0 \in \mathbb{Q}^m$ and $H = \{x : c^\top \cdot x - v = 0\}$, we first convert A to real Jordan normal form as described in Section 2.3.1 to obtain $J = Q^{-1}AQ$. We then perform a coordinate transform on x_0 and H to obtain $H' = \{x : c^\top Qx - v = 0\}$ and $x_0' = Q^{-1}x_0$. The original problem is now equivalent to pseudo-reachability of H' from x_0' under J.

Next, we remove dimensions from $x_0', c^\top Q$ and J that do not correspond to relevant blocks and determine the relevant spectral radius $\rho_H(J)$ of J. If $\rho_H(J) = 1$ then H' is pseudo-reachable by Lemma 4.19(1). If $\rho_H(J) < 1$, then by Lemma 4.19(2), H' is pseudo-reachable if and only if $\mathbf{0} \in H'$ or $x_0', Jx_0', \ldots, J^N x_0'$ hits H', where N is the computable bound in the Lemma.

Finally, we consider the case where $\rho_H(J) > 1$. Let $J_1, \ldots, J_t$ be the blocks of J with $\rho(J) = \rho_H(J)$ and $c^1, \ldots, c^t$, $x_0^1, \ldots, x_0^t$ be the corresponding coordinates of $c^\top Q$ and x_0', respectively. Finally, compute the value of $\liminf_{n\to\infty} \left|\sum_{i=1}^t g^i(n)x_n^i\right|$ using Lemma 4.22 and use Lemma 4.21 to either immediately conclude pseudo-reachability or to compute the bound N and determine reachability by checking if $x_0', Jx_0', \ldots, J^N x_0'$ hits H'.

□

A positive example of hyperplane pseudo-reachability is shown in Figure 4.1.

4.6 Computing the limit inferior of an exponential-polynomial sum

We now prove a generalization of Lemma 4.22. Let $\lambda_1, \ldots, \lambda_m$ be algebraic numbers of modulus 1 and let $p_1, \ldots, p_m$ be polynomials with algebraic coefficients. Let n range over the natural numbers. We show how to effectively determine the value of $\liminf_{n\to\infty} |\sum_{j=1}^m p_j(1/n)\lambda_j^n|$. Moreover, if the value is strictly greater than 0, we show we can find an explicit bound Δ and $N \in \mathbb{N}$ such that for all $n > N$, we have $|\sum_{j=1}^m p_j(1/n)\lambda_j^n| > \Delta$. Lemma 4.22 follows as a special case.

4.6.1 Polynomials to constants

First we show that we can replace the polynomials by their constant terms.

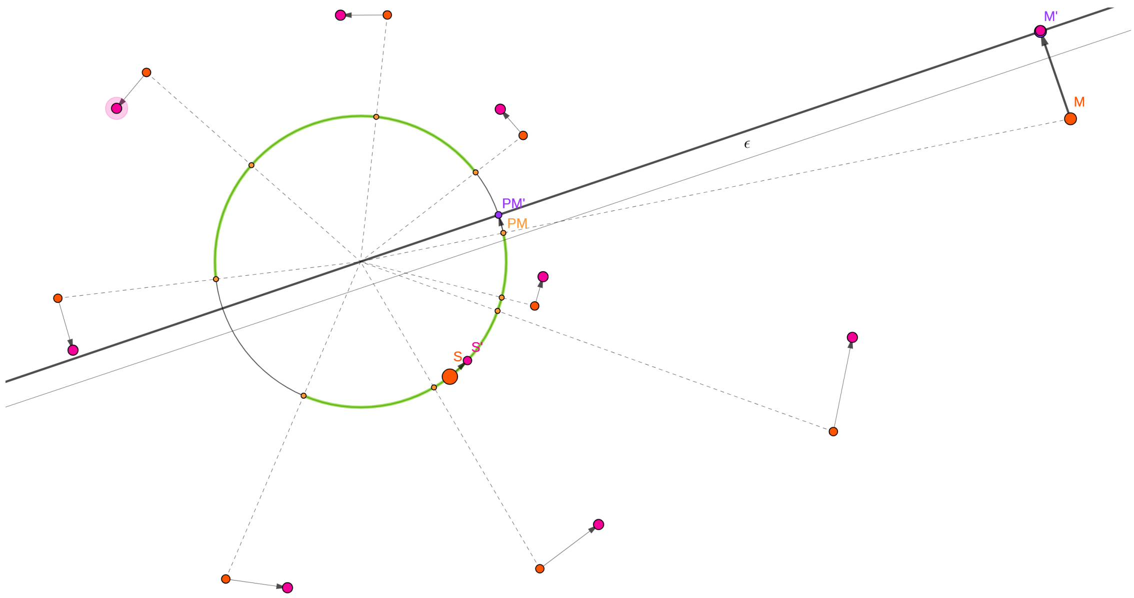

Figure 4.1: By normalizing the required perturbation $M \to M'$, we see that the required perturbation $PM \to PM'$ can be implemented via an initial perturbation $S \to S'$.

Lemma 4.24. *Let $\lambda_1, \ldots, \lambda_m$ be complex numbers of modulus 1. Let $p_1, \ldots, p_m$ be polynomials (with algebraic coefficients) with constant terms $c_1, \ldots, c_m$ respectively. Then*

$$\liminf_{n\to\infty} \left| \sum_{j=1}^{m} p_j(1/n)\lambda_j^n \right| = \liminf_{n\to\infty} \left| \sum_{j=1}^{m} c_j\lambda_j^n \right|.$$

Proof. We can write $p_j(1/n)$ as $c_j + \sum_{i=1}^{d_j} c_{(j,i)}\frac{1}{n^i}$, where c_j is the constant term, $c_{(j,i)}$ are the other coefficients, and d_j is the degree. Define $A_i = \sum_{i=1}^{d_j} |c_{(j,i)}|$ and observe that

$$|p_j(1/n) - c_j| < |\sum_{i=1}^{d_j} c_{(j,i)}\frac{1}{n^i}| < \frac{\sum_{i=1}^{d_j} |c_{(j,i)}|}{n} = \frac{A_j}{n}$$

Thus for any ϵ, setting $N_j(\epsilon) = \lceil A_j/\epsilon \rceil$ ensures that

$$n > N_j(\epsilon) \implies |p_j(1/n) - c_j| < \epsilon.$$

Now let S_n be defined as $|\sum_{j=1}^{m} c_j\lambda_j^n|$.

Observe that

$$|\sum_{j=1}^{m} p_j(1/n)\lambda_j^n| = |\sum_{j=1}^{m}(c_j + p_j(1/n) - c_j)\lambda_j^n|.$$

Thus

$$S_n - |\sum_{j=1}^{m}(p_j(1/n) - c_j)\lambda_j^n| \le |\sum_{j=1}^{m} p_j(1/n)\lambda_j^n| \le S_n + |\sum_{j=1}^{m}(p_j(1/n) - c_j)\lambda_j^n|$$

and so

$$S_n - \sum_{j=1}^{m} |(p_j(1/n) - c_j)\lambda_j^n| \leq |\sum_{j=1}^{m} p_j(1/n)\lambda_j^n| \leq S_n + \sum_{j=1}^{m} |(p_j(1/n) - c_j)\lambda_j^n|,$$

by elementary properties of sums of absolute values. Because the λ_js have absolute value 1, we can reduce the proposition above to

$$S_n - \sum_{j=1}^{m} |(p_j(1/n) - c_j)| \leq |\sum_{j=1}^{m} p_j(1/n)\lambda_j^n| \leq S_n + \sum_{j=1}^{m} |(p_j(1/n) - c_j)|.$$

Now setting $n > N(\epsilon) := \max_{j \in \{1,\ldots,m\}} N_j(\epsilon/m)$, we have $|(p_j(1/n) - c_j)| < \epsilon/m$ for all j, which gives us

$$S_n - \epsilon \leq \left| \sum_{j=1}^{m} p_j(1/n)\lambda_j^n \right| \leq S_n + \epsilon$$

Taking the limit inferior of each term gives us the desired result. □

4.6.2 Diophantine approximation

Kronecker's theorem (see section 2.6) enables us to compute the lim inf by minimizing a function over a compact algebraic set:

Theorem 4.25. *Let $\lambda_1, \ldots, \lambda_m$ be complex numbers of modulus 1. Let $p_1, \ldots, p_m$ be polynomials (with algebraic coefficients) with constant terms $c_1, \ldots, c_m$ respectively. Let $\mathbf{z} = (z_1, \ldots, z_m)$ and $\mathbf{c} = (c_1, \ldots, c_m)$. We have that*

$$\liminf_{n\to\infty} \left| \sum_{j=1}^{m} p_j(1/n)\lambda_j^n \right| = \liminf_{n\to\infty} \left| \sum_{j=1}^{m} c_j\lambda_j^n \right| = \inf_{\mathbf{z}\in\mathcal{T}} |\mathbf{c}^\top \cdot \mathbf{z}| = \min_{\mathbf{z}\in\mathcal{T}} |\mathbf{c}^\top \cdot \mathbf{z}|,$$

where $\mathcal{T}$ is the algebraic set which is the closure of $\{(\lambda_1^n, \ldots, \lambda_m^n) : n \in \mathbb{N}\}$.

Proof. The first equality follows from Lemma 4.24 and the second follows from Theorem 2.16. The third equality holds because the function $\mathbf{z} \mapsto |\mathbf{c}^\top \cdot \mathbf{z}|$ is continuous and T is compact. □

4.6.3 Computing with the theory of the reals

Now, since $\mathcal{T}$ is an algebraic set, the minimum $\min_{\mathbf{z}\in\mathcal{T}} |\mathbf{c}^\top \cdot \mathbf{z}|$ can be expressed in the theory of reals with addition and multiplication (omitting the encoding of absolute values):

$$\exists \mathbf{z} \in \mathcal{T}.v = |\mathbf{c}^\top \cdot \mathbf{z}| \wedge \forall \mathbf{z}' \in T.v \leq |\mathbf{c}^\top \cdot \mathbf{z}'|$$

Therefore, by Tarski's theorem [82, 5, 25], we can characterize the unique v that attains the minimum.

Suppose the minimum v is some number $B > 0$. In this case, we need a bound $\Delta \in \mathbb{R}$ and $N \in \mathbb{N}$ such that $|\sum_{j=1}^{m} p_j(1/n)\lambda_j^n| > \Delta$ for all $n > N$. By emulating the proof of Lemma 4.24, we can find a bound N such that for all $n > N$, we have $|\sum_{j=1}^{m} p_j(1/n)\lambda_j^n| > B/2$. The required bounds are $\Delta = B/2$ and this N.

This concludes the proof of Lemma 4.22 and therefore also Theorem 4.23.

4.7 Conclusion

In this chapter, we have analysed reachability problems from a more topological perspective, with surprisingly different results from the exact versions of the problems. Our results can also be applied to dynamical systems and control theory. Conley [26] formulated the fundamental theorem of dynamical systems: the iteration of any continuous, possibly non-linear, map on a compact metric space decomposes the space into a chain-recurrent part (the pseudo-orbit analogue of a period orbit) and a gradient-like part. Our results imply that deciding if a state is chain recurrent is decidable for linear systems.

In linear systems theory, *controllability* is a fundamental property of linear systems [89]. Controllability states that the system can be controlled from any point to any other point. However, this may require unboundedly large control inputs. A pseudo-orbit can be seen as a stronger notion, where we ask if the dynamics can be controlled from a starting point to an ending point no matter how small the control input is: if a state belongs to the pseudo-orbit, then for every ϵ, there is a sequence of control inputs each bounded in norm by ϵ that steers the system to that state.

A note on limitations of the approach in this chapter: our techniques take advantage of the fact that a hyperplane is defined by a single equation, and thus we can analyse the kind of perturbation necessary to satisfy this equation. Unfortunately, this quickly becomes intractable for even two equations, such as determining if we can pseudo-reach the intersection of two hyperplanes, or any more complex unbounded semialgebraic set. In the next chapter, we will consider a different approach that abstracts away from the details of the perturbations, which allows us to extend our results to more complex sets.

Chapter 5

Pseudo-Reachability for Diagonalisable Systems

5.1 Introduction

The *pseudo-reachability problem for affine dynamical systems* generalizes the problem we considered in the previous chapter: it asks, given a starting point s, affine dynamics $x \mapsto Mx + b$, and semialgebraic target S, whether for all $\epsilon > 0$ there exists an ϵ-pseudo-orbit that reaches S. In other words, for every positive value of ϵ (provided by the environment), does there exist a sequence $\langle x_0 = s, x_1, x_2 \ldots x_N \rangle$ such that $||Mx_n + b - x_{n+1}||_\infty \leq \epsilon$ for all $0 \leq n < N$ and $x_N \in S$?

As we mentioned earlier, already for a target S that is an intersection of two halfspaces, deciding pseudo-reachability becomes difficult because given $\epsilon > 0$, there is no compact description of all ϵ-pseudo-orbits that reach each of the halfspaces that would allow one to check whether there exists a single ϵ-pseudo-orbit that reaches both halfspaces at the same time.

In this chapter, we develop a completely different "logical" approach to show the decidability of the pseudo-reachability problem for diagonalisable systems with arbitrary semialgebraic targets.

5.1.1 High-level proof sketch of our approach

Our solution to the diagonalisable pseudo-reachability problem can be summarised as follows. Let $\widetilde{\mathcal{O}}_\epsilon(n)$ denote the set of all points that are reachable exactly at time n via an ϵ-pseudo-orbit. The pseudo-reachability problem then consists in checking whether the sentence $\Phi \coloneqq \forall \epsilon. \exists n \in \mathbb{N} : \widetilde{\mathcal{O}}_\epsilon(n) \cap S \neq \emptyset$ is true. In this form, Φ is not amenable to application of logical methods as it involves both integer and real-valued variables, in addition to exponentiation with a complex base (coming from non-real eigenvalues of M). We therefore first move to the continuous domain and construct an abstraction $\mathcal{A}_\epsilon(t)$ for $t \in \mathbb{R}^{\geq 0}$ that is definable in $\mathbb{R}_{\exp}$ such that $\mathcal{A}_\epsilon(n) \supseteq \widetilde{\mathcal{O}}_\epsilon(n)$ for all $n \in \mathbb{N}$. We then investigate the values of ϵ and t that make $\Psi(\epsilon, t) \coloneqq \mathcal{A}_\epsilon(t) \cap S \neq \emptyset$ true. We show that by the o-minimality of $\mathbb{R}_{\exp}$, either for every $\epsilon > 0$ there exists T such that for all $t > T$, $\Psi(\epsilon, t)$ holds, or the pseudo-reachability problem is equivalent to a finite-horizon reachability problem that is easily solvable. In the former case, it follows that for every $\epsilon > 0$, $\Psi(\epsilon, n)$ holds for all sufficiently large integer values n, thus establishing a bridge back to the discrete setting. We conclude by showing that in this case, S is pseudo-reachable. Intuitively, the idea is to use the universal quantification over ϵ to argue that if S can be reached using an $\epsilon/2$-abstraction at all but finitely many time steps, then it can be reached by an ϵ-pseudo-orbit, in fact at infinitely many possible time steps. The importance of the universal quantification is also illustrated by the following hardness result. For any fixed $\epsilon > 0$, it is decidable whether $\exists n \in \mathbb{N} : \mathcal{A}_\epsilon(n) \cap S \neq \emptyset$ holds, whereas the ϵ-pseudo-reachability problem of determining whether $\exists n \in \mathbb{N} : \widetilde{\mathcal{O}}_\epsilon(n) \cap S \neq \emptyset$ holds is hard with respect to (a hard subclass of) the Skolem problem, as shown in section 5.3.

The approach outlined above can be adapted to solve a few other related problems about linear dynamical systems. An example would be the *robust reachability problem* considered by Akshay et al. in [1]: given an LDS $\langle M, s\rangle$ and a semialgebraic target S, decide whether for all $\epsilon > 0$ there exists a point s' in the ϵ-neighbourhood of s whose orbit reaches S. This problem can be thought of as a modification of the pseudo-reachability problem where only one perturbation is allowed at the very beginning. Due to this simplification, we are able to show, in subsection 5.5.1, full decidability (that is, without the restriction to diagonalisable systems) of the robust reachability problem. Finally, because the first step of our solution is to translate the problem into the continuous domain, the continuous versions of both the pseudo-reachability problem (discussed in section 5.4) and the robust reachability problem (discussed in subsection 5.5.2) can be handled using the same approach, arguably more naturally. For the former, because we proceed by reducing the pseudo-reachability problem to the bounded-time reachability problem, the decidability result assumes Schanuel's conjecture.

5.2 Decidability for discrete-time diagonalisable systems

In this section we prove our main result: the decidability of the pseudo-reachability problem for discrete-time diagonalisable *affine dynamical systems*, which are a generalisation of LDS.
The reason we consider affine systems is that the well-known homogenisation trick (increasing the dimension by one and adding a coordinate that is always equal to 1) used for reducing the classical reachability problem for affine systems to the reachability problem for LDS doesn't work for the pseudo-reachability problem: when perturbations are allowed, one cannot force a coordinate to remain constant. Hence affine systems require separate treatment.

Problem 5.1 (pseudo-reachability)**.** *Let $M \in \mathbb{Q}^{m\times m}$ be an update matrix, $s \in \mathbb{Q}^m$ be a starting point, $b \in \mathbb{Q}^m$ be an affine term and $S \subseteq \mathbb{R}^m$ be a semialgebraic target set. A sequence $\langle x_0 = s, x_1, x_2 \ldots\rangle$ is an ϵ-pseudo-orbit of s if $||Mx_n + b - x_{n+1}||_\infty \leq \epsilon$ for all n. The pseudo-reachability problem asks: given M, b, s and S, decide whether for each $\epsilon > 0$ there exists an ϵ-pseudo-orbit of s that reaches the set S.*

Let $\widetilde{\mathcal{O}}_\epsilon(n)$ denote the set of all points that are reachable via an ϵ-pseudo-orbit of s under the map $x \mapsto Mx + b$ at time n. Since $\widetilde{\mathcal{O}}_\epsilon(0) = \{s\}$ and $\widetilde{\mathcal{O}}_\epsilon(n+1) = M\widetilde{\mathcal{O}}_\epsilon(n) + b + \epsilon B(\mathbf{0}, 1)$, by induction we can show that $\widetilde{\mathcal{O}}_\epsilon(n) = M^n s + \sum_{i=0}^{n-1} M^i b + \epsilon \sum_{i=0}^{n-1} M^i B(\mathbf{0}, 1)$. The pseudo-reachability problem is then equivalent to determining the truth of $\forall \epsilon.\, \exists n : \widetilde{\mathcal{O}}_\epsilon(n) \cap S \neq \emptyset$. Here $B(\mathbf{0}, 1)$ can be viewed as a set of "control inputs", and the pseudo-reachability problem can be viewed as the problem of determining whether S can be reached using arbitrarily small control inputs. The next lemma shows that we can in fact, choose any reasonable control set.

Lemma 5.2 (Invariance under change of the control set)**.** *Let $\mathcal{B} \subseteq \mathbb{R}^m$ be a bounded set containing an open ball around the origin.*

1. *The pseudo-reachability problem as defined above is equivalent to the problem of determining whether*

$$\forall \epsilon.\, \exists n : (M^n s + \sum_{i=0}^{n-1} M^i b + \epsilon \sum_{i=0}^{n-1} M^i \mathcal{B}) \cap S \neq \emptyset.$$

2. *We may assume the matrix M is in real Jordan form.*

Proof. Since $\mathcal{B}$ is assumed to be bounded and to contain an open neighbourhood around the origin, there must exist constants C_1, C_2 such that $C_1 B(\mathbf{0}, 1) \subseteq \mathcal{B} \subseteq C_2 B(\mathbf{0}, 1)$. Hence

$$C_1 \epsilon \sum_{i=0}^{n-1} M^i B(\mathbf{0}, 1) \subseteq \epsilon \sum_{i=0}^{n-1} M^i \mathcal{B} \subseteq C_2 \epsilon \sum_{i=0}^{n-1} M^i B(\mathbf{0}, 1).$$

The proof of (1) then follows from the fact that ϵ is universally quantified: one can simulate (i) an ϵ-pseudo-orbit with control set $\mathcal{B}$ using a $C_2\epsilon$-pseudo-orbit with control set $B(\mathbf{0}, 1)$ and (ii) an ϵ-pseudo-orbit with control set $B(\mathbf{0}, 1)$ using a ϵ/C_1-pseudo-orbit with control set $\mathcal{B}$. The proof of (2) follows from observing that multiplying $B(\mathbf{0}, 1)$ by an invertible change of basis matrix results in a bounded control set containing a neighbourhood around $\mathbf{0}$. □

Observe that the change of the control set described above is not applicable when ϵ is fixed, as in the ϵ-pseudo-reachability problem discussed in section 5.3.

5.2.1 A closed form for $\widetilde{\mathcal{O}}_\epsilon(n)$

We now use Lemma 5.2 to choose a control set that results in $\widetilde{\mathcal{O}}_\epsilon(n)$ with a convenient first-order closed form: observe that the naïve formulation above involves the term $\sum_{i=0}^{n-1} M^i B(\mathbf{0}, 1)$ which is not "first-order".

Assume M is diagonalisable and in real Jordan form:

$M = \operatorname{diag}(\Lambda_1, \ldots, \Lambda_k, \rho_{k+1}, \ldots, \rho_d)$. That is, M consists of d blocks, the first k of which have dimension 2×2 and a pair of non-real conjugate eigenvalues, whereas the remaining blocks are 1×1 and real. Write ρ_j for the spectral radius of the jth block. We can factor M into a "scaling" and a "rotation" as $M = DR$ where $D = \operatorname{diag}(\rho_1, \rho_1, \ldots, \rho_k, \rho_k, \rho_{k+1}, \rho_{k+2}, \ldots, \rho_d)$ is diagonal and R is a block-diagonal matrix that consists of blocks that are either 2×2 rotation matrices or 1×1 and equal to $\left[\pm 1\right]$. Hereafter we will be using the convenient "rotation-invariant" control set

$$\mathcal{B} = \prod_{j=1}^{k} B((0,0), 1) \times \prod_{j=k+1}^{d} [-1, 1] = \prod_{j=1}^{d} B(\mathbf{0}, 1)$$

where $B((0,0), 1)$ is the unit disc. Observe that $\mathcal{B}$ is a product of ℓ_2-balls that matches the block structure of M. It follows that $R\mathcal{B} = \mathcal{B}$ and hence

$$\widetilde{\mathcal{O}}_\epsilon(n) = D^n R^n s + \sum_{i=0}^{n-1} M^i b + \epsilon \sum_{i=0}^{n-1} D^i R^i \mathcal{B} = D^n R^n s + \sum_{i=0}^{n-1} M^i b + \epsilon \mathcal{B}(n)$$

where $\mathcal{B}(n) = \sum_{i=0}^{n-1} D^i \mathcal{B}$. We then have

$$\mathcal{B}(n) = \sum_{i=0}^{n-1} D^i \prod_{j=1}^{d} B(\mathbf{0}, 1) = \sum_{i=0}^{n-1} \prod_{j=1}^{d} B(\mathbf{0}, \rho_j^i) = \prod_{j=0}^{d} B(\mathbf{0}, \sum_{i=0}^{n-1} \rho_j^i).$$

Geometrically, the idea is that a 2×2 or a 1×1 block of D maps an origin-centred disc (which corresponds to a symmetric interval in 1D) to an origin-centred disc, and a set-sum of such discs is again an origin-centred disc. Note that our ability to reason in this way crucially depends on the fact that M is diagonalisable. Finally, since $\sum_{i=0}^{n-1} \rho_j^i$ is either $\frac{\rho_j^n - 1}{\rho_j - 1}$ or $n\rho_j$, we can write $\mathcal{B}(n) = \{z : \varphi(z, n, \rho_1^n, \ldots, \rho_d^n)\}$, where φ is a semialgebraic predicate.

We can apply the blockwise summation technique, distinguishing between the cases where the spectral radius of the block is 1 or different from 1, to the term $\sum_{i=0}^{n-1} M^i b$ to obtain the closed form $\sum_{i=0}^{n-1} M^i b = D^n R^n x' + cn + d$, where x', c and d only depend on M and b. We then fold s and x' into a new, fictive starting point x to obtain the final closed form

$$\widetilde{\mathcal{O}}_\epsilon(n) = D^n R^n x + cn + d + \epsilon \mathcal{B}(n).$$

In order to solve the pseudo-reachability problem, we henceforth consider the problem of determining the truth of the sentence $\forall \epsilon > 0. \exists n : (D^n R^n x + cn + d + \epsilon \mathcal{B}(n)) \cap S \neq 0$, where all the input vectors and matrices have real algebraic entries.

5.2.2 Passing to the abstraction

The expression for $\widetilde{\mathcal{O}}_\epsilon(n)$ contains the term $D^n R^n x$, which is the last obstacle to obtaining an expression which we can attack using known results about theories of real numbers. To address this issue we resort to abstracting $\widetilde{\mathcal{O}}_\epsilon(n)$. Let

$$\mathcal{T} := \mathrm{cl}\left(\{R^n x : n \in \mathbb{N}\}\right) \text{ and } \mathcal{A}_\epsilon(n) := D^n \mathcal{T} + cn + d + \epsilon \mathcal{B}(n)$$

where $\mathcal{T}$ is the closure of the orbit of x under R, and is semialgebraic and effectively computable by Corollary 2.17. Moreover, recall that by Kronecker's theorem for every $z \in \mathcal{T}$ and $\epsilon > 0$ there exist infinitely many integers $0 < n_1 < n_2 < \ldots$ such that $||R^{n_i} x - z|| < \epsilon$ for all i.

Here $\mathcal{A}_\epsilon(n)$ acts as an abstraction of $\widetilde{\mathcal{O}}_\epsilon(n)$. In particular, for all $\epsilon > 0$ and $n \in \mathbb{N}$ we have $\mathcal{A}_\epsilon(n) \supseteq \widetilde{\mathcal{O}}_\epsilon(n)$. Observe that $\mathcal{A}_\epsilon(n) = \{z : \varphi(z, \epsilon, n, \rho_1^n, \ldots, \rho_d^n)\}$ for a semialgebraic predicate φ. Viewing $\mathcal{A}_\epsilon(n)$ as a proxy for $\widetilde{\mathcal{O}}_\epsilon(n)$, we arrive at the following dichotomy.

Lemma 5.3 (Dichotomy lemma)**.** *Either*

1. *for every $\epsilon > 0$ there exists N_ϵ such that for all $n > N_\epsilon$, $\mathcal{A}_\epsilon(n)$ intersects S, or*
2. *there exist N and $\epsilon > 0$, both effectively computable, such that $\mathcal{A}_\epsilon(n)$ does not intersect S for all $n > N$.*

Moreover, it can be effectively determined which case holds.

Proof. First we show that the dichotomy holds, putting the issues of effectiveness aside. Let

$$\Phi(\epsilon, n) = \bigvee_{\alpha \in A} \bigwedge_{\beta \in B} p_{\alpha,\beta}(\epsilon, n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0$$

be a quantifier-free formula equivalent to $\mathcal{A}_\epsilon(n) \cap S \neq \emptyset$. Such $\Phi(\epsilon, n)$ must exist because $\mathcal{A}_\epsilon(n)$ is semialgebraic with parameters from $\{\epsilon, n, \rho_1^n, \dots, \rho_d^n\}$ and by the Tarski-Seidenberg theorem, each such set can be described using a quantifier-free formula of the form given above. Suppose Case 1 does not hold. Then there exists a particular $\epsilon > 0$ such that $\Phi(\epsilon, n)$ does not hold for arbitrarily large n. Treating n as a continuous parameter, consider the set $\{n \in \mathbb{R}_{\geq 0} : \Phi(\epsilon, n) \text{ does not hold}\}$. By o-minimality of $\mathbb{R}_{\exp}$ this set is a finite union of intervals and and by the assumption that Case 1 does not hold, it contains arbitrarily large integers. Hence it must contain all integers in (N, ∞) for some $N \in \mathbb{N}$. That is, for all $n > N$ the formula $\Phi(\epsilon, n)$ does not hold.

Effectiveness. We now address the issues of effectiveness. Consider the formula

$$\Psi(\epsilon) = \exists N_\epsilon. \forall n > N_\epsilon : \Phi(\epsilon, n).$$

We show that $\Psi(\epsilon)$ is equivalent to a formula $\psi(\epsilon)$ in the language of $\mathbb{R}_0$. To determine which case holds it then remains to determine the truth value of the sentence $\forall \epsilon : \psi(\epsilon)$.

By the o-minimality argument above, given $\epsilon > 0$, each $p_{\alpha,\beta}(\epsilon, n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0$ either holds for finitely many integer values of n or holds for all sufficiently large integer values n. By elementary considerations it follows that $\Psi(\epsilon)$ is equivalent to

$$\bigvee_{\alpha \in A} \bigwedge_{\beta \in B} \exists N_\epsilon. \forall n > N_\epsilon : p_{\alpha,\beta}(\epsilon, n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0.$$

Hence it suffices to show how to construct a formula $\psi(\epsilon)$ in the language of $\mathbb{R}_0$ that is equivalent to $\exists N_\epsilon. \forall n > N_\epsilon : p_{\alpha,\beta}(\epsilon, n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0$. For each $\epsilon > 0$, the formula first tests if $p_{\alpha,\beta}(\epsilon)$ (as a polynomial in $d+1$ remaining variables) is identically zero. If yes, then $\varphi(\epsilon)$ is true or false depending only on $\bowtie_{\alpha,\beta}$. Otherwise, write $p_{\alpha,\beta}(\epsilon, n, \rho_1^n, \dots, \rho_d^n) = \sum_{i=1}^k q_i(\epsilon, n)R_i^n$ where $q_i(\epsilon)$ is not identically zero for all i and $R_1 > \cdots > R_k > 0$ are real algebraic numbers of the form $\rho_1^{p_1} \cdots \rho_d^{p_d}$ for $p_1, \dots, p_d \in \mathbb{N}$. Since $|q_1(\epsilon, n)R_1^n| > \left|\sum_{i=2}^k q_i(\epsilon, n)R_i^n\right|$ for sufficiently large n, whether $p_{\alpha,\beta}(\epsilon, n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0$ holds for sufficiently large n depends only on $q_1(\epsilon, n)$. Hence we can choose $\psi(\epsilon)$ to be $\lim_{n\to\infty} q_1(\epsilon) \bowtie_{\alpha,\beta} 0$, which amounts to a sign condition on the coefficients of $q_1(\epsilon, n)$.

Computing N**.** Finally, we show that in Case 2, the value N can be effectively computed. To this end, by repeatedly trying smaller and smaller values of ϵ first compute a rational $e > 0$ such that $\Psi(e)$ (equivalently, $\psi(e)$) does not hold. To be able to compute N it then suffices to compute, for a particular (α, β), a value $N_{\alpha,\beta}$ such that $p_{\alpha,\beta}(e, n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0$ does not hold for all $n > N_{\alpha,\beta}$, assuming that it does not hold for sufficiently large n. We can then take N to be the maximum of $N_{\alpha,\beta}$ over $(\alpha, \beta) \in A \times B$.

To compute $N_{\alpha,\beta}$, consider $p := p_{\alpha,\beta}(e)$. Assuming it is not identically zero (otherwise we can choose $N_{\alpha,\beta}$ to be any positive integer), write $p(n, \rho_1^n, \dots, \rho_d^n) = \sum_{i=1}^k q_i(n)R_i^n$ where q_i is not identically zero for all i and $R_1 > \cdots > R_k > 0$ are real algebraic. Since $p_{\alpha,\beta}(e, n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0$ and hence $p(n, \rho_1^n, \dots, \rho_d^n) \bowtie_{\alpha,\beta} 0$ do not hold for sufficiently large n, it must be the case that $q_1(n) \bowtie_{\alpha,\beta} 0$ does not hold for sufficiently large n. Hence it remains to choose $N_{\alpha,\beta}$ large enough so that for all $n > N_{\alpha,\beta}$, $|q_1(n)R_1^n|$ dominates $\left|\sum_{i=2}^k q_i(n)R_i^n\right|$. □

5.2.3 From the abstraction back to ϵ-pseudo-orbits

In this section we consider the relationship between the two cases of Lemma 5.3 and our original pseudo-reachability problem.

We start with Case 2. Observe that $\mathcal{O}_\epsilon(n) \subseteq \mathcal{A}_\epsilon(n)$ for every $n \in \mathbb{N}$ and $\epsilon > 0$. Therefore, when Case 2 holds, for any $n > N$ and $\epsilon > 0$ the target set cannot be reached by $\mathcal{O}_\epsilon(n)$. It remains to check pseudo-reachability at time steps $n \leq N$. Thus we know that S is pseudo-reachable if and only

$$\forall \epsilon. \exists n \leq N : (M^n x + cn + d + \epsilon\mathcal{B}(n)) \cap S \neq \emptyset.$$

Let $\overline{S}$ denote the topological closure of S. We show that the statement above is equivalent to $\exists n \leq N : M^n x + cn + d \in \overline{S}$. Observe that if for all $n \leq N$ the point $M^n x + cn + d$ is not in $\overline{S}$, then by compactness the smallest distance from $\{M^n x + cn + d \mid n \leq N\}$ to $\overline{S}$ is positive and hence for sufficiently small ϵ the target S cannot be ϵ-pseudo-reached within the first N steps. Conversely, if $M^n x + cn + d \in \overline{S}$ for some $n \leq N$, then because $\mathcal{B}(n)$ is full dimensional and contains $\mathbf{0}$ in its interior,

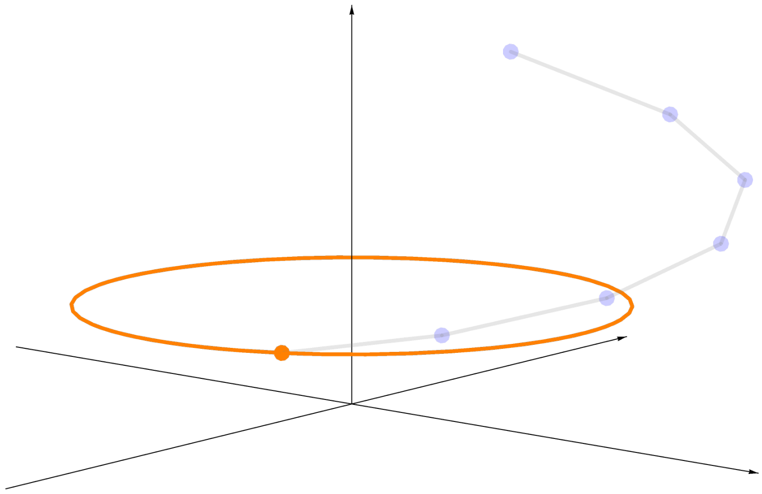

Figure 5.1: The orange circle represents the abstraction $\mathcal{A}_\epsilon(n) := D^n\mathcal{T}+cn+d+\epsilon\mathcal{B}(n)$ at time $n = 1$ (say). The sequence of points represents a particular localisation.

it follows that $(M^n x + cn + d + \epsilon\mathcal{B}(n)) \cap S \neq \emptyset$ for all $\epsilon > 0$. Therefore, in Case 2 pseudo-reachability can be decided by simply checking if $\{M^n x + cn + d \mid n \leq N\}$ reaches $\overline{S}$.

Next we will show that S is pseudo-reachable if Case 1 holds. Given $z \in \mathcal{T}$, we define a "localisation" of the abstraction at the point z as $\mathcal{A}_\epsilon(n)(z) := D^n z + cn + d + \epsilon\mathcal{B}(n)$. Observe that $\mathcal{A}_\epsilon(n) = \{\mathcal{A}_\epsilon(n)(z) : z \in \mathcal{T}\}$. This definition of a localisation will allow us to select a "concrete trajectory" from the set of all possible (abstract) trajectories. This is illustrated in Figure 5.1.

Fix $\epsilon > 0$ and let $T_n := \{z \in \mathcal{T} : \mathcal{A}_\epsilon(n)(z) \text{ intersects } S\}$. The next lemma implies that the sequence T_n must tend towards a limiting shape; i.e. it cannot "jump around" forever.

Lemma 5.4. *Let $T_n = \{z : \varphi(z, n, \rho_1^n, \ldots, \rho_d^n)\}$, where φ is a semialgebraic predicate and $\rho_1, \ldots, \rho_d$ are real algebraic, be a family of non-empty sets contained in a compact set $\mathcal{T}$. There exists a non-empty limiting set $L \subseteq \mathcal{T}$ to which the sequence T_n converges as $n \to \infty$, in the following sense.*

1. *For every $\epsilon > 0$, there exists N such that for all $n > N$, $T_n \subseteq L + B(\mathbf{0}, \epsilon)$.*

2. *For all $z \in L$ and $\epsilon > 0$ there exists N such that for all $n > N$, $z + B(\mathbf{0}, \epsilon)$ intersects T_n.*

Proof. Write $\varphi(z, n, \rho_1^n, \ldots, \rho_d^n) = \bigvee_{\alpha \in A} \bigwedge_{\beta \in B} p_{\alpha,\beta}(z, n, \rho_1^n, \ldots, \rho_d^n) \bowtie_{\alpha,\beta} 0$. We can define the sequence $\langle T_t \mid t \in \mathbb{R} \rangle$ as $T_t = \{z : \varphi(z, t, \rho_1^t, \ldots, \rho_d^t)\}$. Let $L = \{x : \liminf d(x, T_t) = 0\}$ where $d(x, T_t)$ denotes the shortest Euclidean distance from x to a point in T_t.

We prove the first claim by contradiction. Suppose there exists $\epsilon > 0$ such that at infinitely many unbounded time steps $t_1 < t_2 < \ldots$ there are points $z_1, z_2, \ldots$ such that $z_i \in T_i$ but $z_i \notin L + B(\mathbf{0}, \epsilon)$. Then the sequence z_i must have an accumulation point z in $\mathcal{T} \setminus L$. But z will also satisfy $\liminf d(z, T_t) = 0$ and hence $z \in L$, a contradiction.

We prove the second claim using o-minimality of $\mathbb{R}_{\exp}$. Fix $z \in L$ and $\epsilon > 0$ and consider the set $Z = \{t \in \mathbb{R} : z + B(\mathbf{0}, \epsilon) \text{ intersects } T_t\}$. The set Z is o-minimal, and since $z \in L$, it is unbounded from above. Hence it must contain an interval of the form (N, ∞), which implies the desired result. □

One can also show that the set L described above is in fact semialgebraic, but this is not necessary for our arguments. We are now ready to show that S is pseudo-reachable if Case 1 of Lemma 5.3 holds.

Lemma 5.5. *If for every $\epsilon > 0$ there exists N_ϵ such that for all $n > N_\epsilon$, $\mathcal{A}_\epsilon(n)$ intersects S then S is pseudo-reachable.*

The main idea of the proof is to use the assumption that $\mathcal{A}_{\epsilon/2}(n)$ intersects S for sufficiently large n to construct an ϵ-pseudo-orbit that hits S. Intuitively, in order to simulate $\mathcal{A}_{\epsilon/2}(n)$ using an ϵ-pseudo-orbit, $\epsilon/2$ of the total control allowance is used to replicate the effect of the control inputs (of size at most $\epsilon/2$, corresponding to the $\frac{\epsilon}{2}\mathcal{B}(n)$ term in the definition of $\mathcal{A}_{\epsilon/2}(n)$) and the remaining $\epsilon/2$ is used to compensate for the abstraction from the starting point x to the set $\mathcal{T}$. In fact, we do not know if one can deduce that S is ϵ-pseudo-reachable from knowing that $\mathcal{A}_\epsilon(n)$ intersects S for sufficiently large n. This illustrates the reason why the pseudo-reachability problem is easier than the ϵ-pseudo-reachability problem; see section 5.3 for a more concrete argument.

Proof. Fix $\epsilon > 0$. We show how to construct an ϵ-pseudo-orbit that hits S. Consider $\mathcal{A}_{\epsilon/2}(n)$. By assumption, there exists N_1 such that for all $n > N_1$, $\mathcal{A}_{\epsilon/2}(n)$ intersects S. We now investigate which localisations of the abstraction are responsible for intersecting S. Apply Lemma 5.4 to the sequence of sets $T_n = \{z \in \mathcal{T} : \mathcal{A}_{\epsilon/2}(n)(z) \text{ intersects } S\}$ to obtain their "limit" L. Fix any $p \in L$.

Let ϵ' be small enough so that $\epsilon' D^n \mathcal{B} \subseteq \frac{\epsilon}{2}\mathcal{B}(n)$ for all $n > 0$. Intuitively, such ϵ' must exist because $D^n\mathcal{B}$ and $D^{n-1}\mathcal{B}$ only differ by at most a constant factor that only depends on the magnitudes $\rho_1, \ldots, \rho_d$ of eigenvalues of M, and we have that $D^{n-1}\mathcal{B} \subseteq \sum_{i=0}^{n-1} D^i\mathcal{B} = \mathcal{B}(n)$. By Lemma 5.4 (b), there exists $N > N_1$ such that for all $n > N$, $p + B(\mathbf{0}, \epsilon'/2)$ intersects T_n. That is, for all $n > N$ there exists $p_n \in \mathcal{T}$ such that $||p - p_n|| < \epsilon'/2$ and $p_n \in T_n$. Equivalently,

$$||p - p_n|| < \epsilon'/2 \text{ and } \mathcal{A}_{\epsilon/2}(n)(p_n) \text{ intersects } S.$$

By Kronecker's theorem there must exist $m > N$ such that $||R^m x - p|| < \epsilon'/2$. Hence we have $||R^m x - p_m|| < \epsilon'$ which implies $R^m x - p_m \in \epsilon'\mathcal{B}$ and hence $D^m(R^m x - p_m) \in \epsilon' D^m \mathcal{B}$. Since by construction of ϵ' we have $\epsilon' D^m\mathcal{B} \subseteq \frac{\epsilon}{2}\mathcal{B}(m)$, it follows that $D^m(R^m x - p_m) \in \frac{\epsilon}{2}\mathcal{B}(m)$ and hence $D^m p_m \in D^m R^m x + \frac{\epsilon}{2}\mathcal{B}(m)$. Therefore,

$$\widetilde{\mathcal{O}}_\epsilon(m) = (D^m R^m x + \frac{\epsilon}{2}\mathcal{B}(m)) + cm + d + \frac{\epsilon}{2}\mathcal{B}(m) \supseteq D^m p_m + cm + d + \frac{\epsilon}{2}\mathcal{B}(m) = \mathcal{A}_{\epsilon/2}(m)(p_m).$$

Since $\mathcal{A}_{\epsilon/2}(m)(p_m)$ intersects S, it then follows that $\widetilde{\mathcal{O}}_\epsilon(m)$ too must intersect S. □

5.2.4 The algorithm

To summarise, the analysis above gives us the following algorithm for determining if S is pseudo-reachable, i.e. if $\forall \epsilon > 0. \exists n : \widetilde{\mathcal{O}}_\epsilon(n) \cap S \neq \emptyset$. Let $\varphi(n, \epsilon)$ be a quantifier-free formula in $\mathbb{R}_{\exp}$ defining the abstraction $\mathcal{A}_\epsilon(n)$. First determine, using the algorithm described in the proof of Lemma 5.3, whether Case 1 or Case 2 holds. If the former holds, then conclude that S is pseudo-reachable. If Case 2 holds, then compute the value of N effectively and check if there exists $n < N$ such that $(M^n x + cn + d) \cap \overline{S} \neq \emptyset$.

5.3 Skolem-hardness of the ϵ-pseudo-reachability problem

In this section we consider the ϵ-pseudo-reachability problem for discrete diagonalisable systems: given diagonalisable M, starting point s, a target set S and fixed $\epsilon > 0$, decide whether there exists n such that $M^n s + \sum_{k=0}^{n-1} M^k B(\mathbf{0}, \epsilon) \cap S \neq 0$. This problem is also known as the reachability problem for linear time-invariant systems [40] with the control set $B(\mathbf{0}, \epsilon)$.

We will reduce a hard subclass of the Skolem problem to our ϵ-pseudo-reachability problem. We will show the hardness of this problem where $B(\mathbf{0}, \epsilon)$ is the unit ℓ_2-ball.

As we have discussed before, the Skolem problem is not known to be decidable for orders $d \geq 5$, even for diagonalisable recurrences. (There is a conditional decidability result for diagonalisable recurrences assuming two strong number-theoretic conjectures [14].) The largest class of sequences for which decidability is known is the MSTV (Mignotte-Shorey-Tijdeman-Vereschagin) class, which consists of all linear recurrence sequences over integers that (i) have at most three dominant roots with respect to the usual (Archimedean) absolute or (ii) have at most two dominant roots with respect to a p-adic absolute value [60]. We consider the Skolem problem for integer sequences whose roots $\rho, \lambda_1, \ldots, \lambda_d$ satisfy $\rho = |\lambda_1| = \cdots = |\lambda_d|$. This class of sequences contains many instances that are not in the MSTV class and hence is a hard subclass of the Skolem problem.

Recall that any linear recurrence sequence can be written as $u_n = c^\top M^n s$ where M is the companion matrix of u_n whose eigenvalues are the roots of u_n. Let u_n be a diagonalisable sequence that belongs to the hard subclass described above, i.e. $u_n = c^\top M^n s$ where $M = \operatorname{diag}(\Lambda_1, \ldots, \Lambda_d, \rho)$ and Λ_i is a 2×2 real Jordan block with $\rho(\Lambda_i) = \rho$ for $1 \leq i \leq d$. We reduce the problem "does u_n have a zero?" to an ϵ-pseudo-reachability problem.

Theorem 5.6 (Skolem reduces to ϵ-pseudo-reachability)**.** *The ϵ-pseudo-reachability problem (for fixed $\epsilon > 0$) is Skolem-hard for diagonalisable sequences.*

Proof. Let u_n be an LRS. Consider the sequence $v_n = u_n^2$. Observe that $v_n = \sum_{i=1}^{L} c_i \Gamma_i^n s_i + C r^n$ where

- $r = \rho^2$,
- Γ_i is a 2×2 real Jordan block with $\rho(\Gamma_i) = r$ for $1 \leq i \leq L$,
- $c_i, s_i \in \mathbb{R}^2$ for $1 \leq i \leq L$, and
- $C > 0$.

The first two statements follow from the fact that the eigenvalues of v_n are products of eigenvalues of u_n. That $C > 0$ can be deduced as follows. Consider $w_n = \sum_{i=1}^{L} c_i \Gamma_i^n s_i$. It only has non-real roots and hence by [70] is infinitely often positive and negative. Hence if C is not positive, then $v_n < 0$ for infinitely many n, which contradicts the fact that $v_n \geq 0$.

Next observe that u_n has a zero iff there exists n such that $v_n \leq 0$. Since we are interested only in the sign of v_n, by scaling v_n by $C(2r)^n$ if necessary we assume that $r \in (0,1)$ and $C = 1$. We will construct an instance of the ϵ-pseudo-reachability problem that is positive if and only if there exists n such that $v_n \leq 0$.

Define

- $A = \operatorname{diag}(\Gamma_1, \ldots, \Gamma_L)$,
- $s = (s_1, \ldots, s_L)$ and $c = (c_1, \ldots, c_L)$,
- $\epsilon = \frac{1-r}{||c||}$, and
- $H = \{z : c^\top \cdot z + 1 \leq 0\}$.

Observe that H is ϵ-pseudo-reachable if and only if $A^n s + \sum_{i=0}^{n-1} A^i B(\mathbf{0}, \epsilon) \cap H \neq 0$ for some n. Since $A^i B(\mathbf{0}, \epsilon) = B(\mathbf{0}, r^i \epsilon)$, we have $\sum_{i=0}^{n-1} A^i B(\mathbf{0}, \epsilon) = B(\mathbf{0}, \frac{1-r^n}{1-r}\epsilon) := \mathcal{B}(n)$ and

$$H \text{ is } \epsilon\text{-pseudo-reachable} \iff \min_{z \in \mathcal{B}(n)} c \cdot (A^n s + z) + 1 \text{ is } \leq 0 \text{ for some } n.$$

We will show that in fact $\min_{z \in \mathcal{B}(n)} c \cdot (A^n s + z) + 1 = v_n$, which will conclude the proof.

$$\begin{aligned}
\min_{z \in \mathcal{B}(n)} c \cdot (A^n s + z) + 1 &= \sum_{i=1}^{L} c_i \Gamma_i^n s_i + 1 + \min_{z \in \mathcal{B}(n)} c \cdot z \\
&= \sum_{i=1}^{L} c_i \Gamma_i^n s_i + 1 - ||c|| \frac{1 - r^n}{1 - r} \epsilon \\
&= \sum_{i=1}^{L} c_i \Gamma_i^n s_i + r^n \\
&= v_n.
\end{aligned}$$

□

5.4 The continuous-time pseudo-reachability problem

In this section we show that the approach we described in section 5.2 for deciding the discrete-time pseudo-reachability problem for diagonalisable systems also works in the continuous setting with one important difference: to handle Case 2 of the dichotomy lemma (exactly the same as Lemma 5.3) we need to solve the *bounded-time reachability problem for continuous-time affine dynamical systems*, which is only known to be decidable assuming Schanuel's conjecture [21].

Let $M = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_k, \rho_{k+1}, \ldots, \rho_d) \in (\mathbb{R} \cap \overline{\mathbb{Q}})^{L \times L}$ be a diagonalisable matrix in real Jordan form, $s \in \mathbb{Q}^L$ be a starting point, $b \in \mathbb{Q}^L$ be an affine term and $S \subseteq \mathbb{R}^L$ be a semialgebraic target set. The trajectory of the system (in the absence of additional control inputs) is given by

$$x(t) = e^{Mt}s + \int_0^t e^{Mh} b \, dh.$$

Intuitively, while in the discrete setting control inputs are applied after each unit of time and thus are represented by a sequence $\langle d_n \mid n \in \mathbb{N} \rangle$, in the continuous setting they are represented by a continuous function $\Delta : \mathbb{R}_{\geq 0} \to \mathbb{R}^L$. Hence an ϵ-pseudo-orbit is defined as a trajectory

$$x(t) = e^{Mt}s + \int_0^t e^{Mh} b \; dh + \int_0^t e^{Mh} \Delta(t-h) \, dh.$$

for some control signal $\Delta : \mathbb{R}_{\geq 0} \to \mathbb{R}^L$ satisfying $||\Delta_\epsilon(t)|| \leq \epsilon$ for all $t \geq 0$. The pseudo-reachability problem is then defined in the same way as before: decide whether for every $\epsilon > 0$ there exists an ϵ-pseudo-orbit that reaches S.

Let $\mathcal{B}$ be the same control set as defined in subsection 5.2.1. For $1 \leq j \leq k$, let $r_j = \mathrm{Re}(\lambda_j)$ and $\omega_j = \mathrm{Im}(\lambda_j)$ where λ_j is a non-real eigenvalue of the block Λ_j. For $k < j \leq d$ let $r_j = \rho_j$ and $\omega_j = 0$. By using essentially the same arguments as in subsection 5.2.1, we can show that the pseudo-reachability problem is equivalent to determining the truth of

$$\forall \epsilon > 0. \, \exists t : (e^{Mt}x + ct + d + \epsilon\mathcal{B}(t)) \cap S \neq \emptyset$$

where x, c, d are L-dimensional vectors and $\mathcal{B}(t) = \{z : \varphi(z, t, e^{r_1 t}, \ldots, e^{r_d t})\}$ for semialgebraic predicate φ. We denote the term $e^{Mt}x + ct + d + \mathcal{B}(t)$ by $\widetilde{\mathcal{O}}_\epsilon(n)$.

To define a convenient abstraction, we again write $e^{Mt} = D(t)R(t)$ where $D(t) \coloneqq \mathrm{diag}(e^{r_1 t}, e^{r_1 t}, \ldots, e^{r_k t}, e^{r_k t}, e^{r_{k+1} t}, e^{r_{k+2} t}, \ldots, e^{r_{k+d} t})$ is diagonal and $R(t)$ is a block diagonal matrix whose blocks are rotation matrices of the form $\begin{bmatrix} \cos(\omega_j t) & -\sin(\omega_i t) \\ \sin(\omega_j t) & \cos(\omega_j t) \end{bmatrix}$ for $1 \leq j \leq k$ and are of the form $\Omega_i = \begin{bmatrix} 1 \end{bmatrix}$ for $k+1 \leq j \leq d$. Just as in the discrete case, we next define

$$\mathcal{T} \coloneqq \mathrm{cl}(\{R(t)x : t \in \mathbb{R}_{\geq 0}\}) \text{ and } \mathcal{A}_\epsilon(t) \coloneqq D(t)\mathcal{T} + ct + d + \epsilon\mathcal{B}(t),$$

where $\mathcal{T}$ is again semialgebraic and effectively computable [21] and $\mathcal{A}_\epsilon$ acts as an abstraction of $\widetilde{\mathcal{O}}_\epsilon$. In particular, for all $\epsilon > 0$ and $t \in \mathbb{R}_{\geq 0}$, we have $\widetilde{\mathcal{O}}_\epsilon(t) \subseteq \mathcal{A}_\epsilon(t)$. Moreover, observe that $\mathcal{A}_\epsilon(t) = \varphi(t, e^{r_1 t}, \ldots, e^{r_d t})$ for a semialgebraic function φ, which is the most important property we need. We use $\mathcal{A}_\epsilon(t)$ in the same way we used $\mathcal{A}_\epsilon(n)$ in the discrete case to arrive at the following dichotomy lemma.

Lemma 5.7. *Either*

1. *for every $\epsilon > 0$ there exists T_ϵ such that for all $t > T_\epsilon$, $\mathcal{A}_\epsilon(t)$ intersects S, or*

2. *there exist T and $\epsilon > 0$, both effectively computable, such that $\mathcal{A}_\epsilon(t)$ does not intersect S for all $t > T$.*

Moreover, it can be effectively determined which case holds.

Proof. The proof is essentially identical to the proof of Lemma 5.3, and in fact even simpler since we may skip the step of converting from a discrete integer variable to a continuous real-valued variable. For clarity we provide it in full.

First we show that the dichotomy holds, putting the issues of effectiveness aside. Let

$$\Phi(\epsilon, t) = \bigvee_{\alpha \in A} \bigwedge_{\beta \in B} p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0$$

be a quantifier-free formula equivalent to $\mathcal{A}_\epsilon(t) \cap S \neq \emptyset$. Suppose Case 1 does not hold. Then there exists a particular $\epsilon > 0$ such that $\Phi(\epsilon, t)$ does not hold for arbitrarily large t. Consider the set $\{t \geq 0 : \Phi(\epsilon, t) \text{ does not hold}\}$. By o-minimality of $\mathbb{R}_{\exp}$ this set is a finite union of intervals and since it contains arbitrarily large real numbers by assumption, it must contain an unbounded interval (T, ∞). That is, for all $t > T$ the formula $\Phi(\epsilon, t)$ does not hold.

Effectiveness: We now address the issue of effectiveness. Consider the formula

$$\Psi(\epsilon) = \exists T_\epsilon . \forall t > T_\epsilon : \Phi(\epsilon, t).$$

We show that $\Psi(\epsilon)$ is equivalent to a formula $\psi(\epsilon)$ in the language of $\mathbb{R}_0$. To determine which case holds it then remains to determine the truth value of the sentence $\forall \epsilon : \psi(\epsilon)$.

By the o-minimality argument above, given $\epsilon > 0$, each of the values of t for which $p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0$ holds is either bounded or contains an unbounded interval. By elementary considerations it follows that $\Psi(\epsilon)$ is equivalent to

$$\bigvee_{\alpha \in A} \bigwedge_{\beta \in B} \exists T_\epsilon . \forall t > T_\epsilon : p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0.$$

Hence it suffices to show how to construct a formula in the language of $\mathbb{R}_0$ that is equivalent to $\exists T_\epsilon . \forall t > T_\epsilon : p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0$. For each $\epsilon > 0$, the formula first tests if $p_{\alpha,\beta}(\epsilon)$ (as a polynomial in $d+1$ remaining variables) is identically zero. If yes, then $\varphi(\epsilon)$ is **true** or **false** depending only on $\bowtie_{\alpha,\beta}$. Otherwise, write $p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) = \sum_{i=1}^{k} q_i(\epsilon, t) R_i^t$ where $q_i(\epsilon)$ is not identically zero for all i and

$R_1 > \cdots > R_k > 0$ are of the form $e^{p_1 r_1 + \ldots + p_d r_d}$ for $p_1, \ldots, p_d \in \mathbb{N}$. Since $|q_1(\epsilon, t)R_1^t| > \left|\sum_{i=2}^k q_i(\epsilon, t)R_i^t\right|$ for sufficiently large t, whether $p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0$ holds for sufficiently large t depends only on $q_1(\epsilon, t)$. Hence we can choose $\psi(\epsilon)$ to be $\lim_{t\to\infty} q_1(\epsilon, t) \bowtie_{\alpha,\beta} 0$, which amounts to a sign condition on the coefficients of $q_1(\epsilon, t)$.

Computing T. Finally, we show that in Case 2, the value T can be effectively computed. To this end, by repeatedly trying smaller and smaller values of ϵ first compute a rational $c > 0$ such that $\Psi(c)$ (equivalently, $\psi(c)$) does not hold. To be able to compute T it then suffices to compute, for a particular (α, β), a value $T_{\alpha,\beta}$ such that $p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0$ does not hold for all $t > T_{\alpha,\beta}$, assuming that it does not hold for sufficiently large t. We can then take T to be the maximum of $T_{\alpha,\beta}$ over $(\alpha, \beta) \in A \times B$.

To compute $T_{\alpha,\beta}$, consider $p := p_{\alpha,\beta}(c)$. Assuming it is not identically zero (otherwise we can choose $T_{\alpha,\beta}$ to be any positive integer), write $p(t, e^{r_1 t}, \ldots, e^{r_d t}) = \sum_{i=1}^k q_i(t)R_i^t$ where q_i is not identically zero for all i and $R_1 > \cdots > R_k > 0$ are of the form $e^{p_1 r_1 + \ldots + p_d r_d}$ for $p_1, \ldots, p_d \in \mathbb{N}$. Since $p_{\alpha,\beta}(\epsilon, t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0$ and hence $p(t, e^{r_1 t}, \ldots, e^{r_d t}) \bowtie_{\alpha,\beta} 0$ do not hold for sufficiently large t, it must be the case that $q_1(t) \bowtie_{\alpha,\beta} 0$ does not hold for sufficiently large t. Hence it remains to choose $T_{\alpha,\beta}$ large enough so that for all $t > T_{\alpha,\beta}$, $|q_1(t)R_1^t|$ dominates $\left|\sum_{i=2}^k q_i(t)R_i^t\right|$. $\square$

We next show pseudo-reachability in Case 1.

Lemma 5.8. *If for every $\epsilon > 0$ there exists T_ϵ such that for all $t > T_\epsilon$, $\mathcal{A}_\epsilon(t)$ intersects S then S is pseudo-reachable.*

Proof. We first define a suitable notion of localisation, in exactly the same way as the discrete case. Given $z \in \mathbb{T}$, let $\mathcal{A}_\epsilon(t)(z) = D(t)z + ct + d + \epsilon\mathcal{B}(t)$.

Fix $\epsilon > 0$. We show how to construct an ϵ-pseudo-orbit that hits S. In fact, this ϵ-pseudo-orbit will hit S at an integer time step m. Consider $\mathcal{A}_{\epsilon/2}(t)$. By assumption, there exists T_1 such that for all $t > T_1$, $\mathcal{A}_{\epsilon/2}(t)$ intersects S. We now investigate which localisations of the abstraction are responsible for intersecting S. Apply Lemma 5.4 to the sets $T_n = \{z \in \mathcal{T} : \mathcal{A}_{\epsilon/2}(n)(z) \text{ intersects } S\}$, $n \in \mathbb{N}$, to obtain their 'limit' L. Fix any $p \in L$.

Let ϵ' be small enough so that for all $n > 0$, $\epsilon' D(n)B(\mathbf{0}, 1) \subseteq \frac{\epsilon}{2}\mathcal{B}(n)$. By Lemma 5.4 (b), there exists $N > T_1$ such that for all integers $n > N$, $p + B(\mathbf{0}, \epsilon'/2)$ intersects T_n. That is, for all $n > N$ there exists $p_n \in \mathcal{T}$ such that $||p - p_n|| < \epsilon'/2$ and $p_n \in T_n$. Equivalently,

$$||p - p_n|| < \epsilon'/2 \text{ and } \mathcal{A}_{\epsilon/2}(n)(p_n) \text{ intersects } S.$$

By Kronecker's theorem there must exist an integer $m > N$ such that $||R(m)x - p|| < \epsilon'/2$. Hence we have $||R(m)x - p_m|| < \epsilon'$ which implies $R(m)x - p_m \in \epsilon' B(\mathbf{0}, 1)$ and hence $D(m)(R(m)x - p_m) \in \epsilon' D(m)B(\mathbf{0}, 1)$. Since by construction of ϵ' we have $\epsilon' D(m)\mathcal{B}(\mathbf{0}, 1) \subseteq \frac{\epsilon}{2}\mathcal{B}(m)$, it follows that $D(m)(R(m)x - p_m) \in \frac{\epsilon}{2}\mathcal{B}(m)$ and hence $D(m)p_m \in D(m)R(m)x + \frac{\epsilon}{2}\mathcal{B}(m)$. Therefore,

$$\begin{aligned}\widetilde{\mathcal{O}}_\epsilon(m) &= \Big(D(m)R(m)x + \frac{\epsilon}{2}\mathcal{B}(m)\Big) + cm + d + \frac{\epsilon}{2}\mathcal{B}(m) \\ &\supseteq D(m)p_m + cm + d + \frac{\epsilon}{2}\mathcal{B}(m) \\ &= \mathcal{A}_{\epsilon/2}(m)(p_m).\end{aligned}$$

Since $\mathcal{A}_{\epsilon/2}(m)(p_m)$ intersects S, it then follows that $\widetilde{\mathcal{O}}_\epsilon(m)$ too must intersect S. □

We thus have reduced the pseudo-reachability problem to that of handling Case 2:

Theorem 5.9. *The continuous-time pseudo-reachability problem reduces to the bounded-time reachability problem for continuous-time affine dynamical systems.*

Proof. Note that the accumulated effect of control sets over a finite-time interval can be made arbitrarily small by choosing small enough ϵ. Therefore, for a bounded-time interval, a semialgebraic set is pseudo-reachable iff its closure is classically reachable. □

Since the bounded-time reachability problem for continuous-time affine dynamical systems can be encoded in $\mathbb{R}_{\exp,\cos\restriction[0,T]}$, we have the following (conditional) decidability result. Note that it is not necessary to use the full Mackintyre-Wilkie decidability algorithm for $\mathbb{R}_{\exp,\cos\restriction[0,T]}$ to solve bounded-time reachability; it can also be done directly using a complex-analytic approach [21] (though still assuming Schanuel's conjecture).

Corollary 5.10. *The continuous-time pseudo-reachability problem for diagonalisable affine dynamical systems is decidable subject to Schanuel's conjecture.*

5.5 Robust Reachability

Robust reachability [1] is a variant of pseudo-reachability, where instead of perturbing the dynamics, one is allowed a single perturbation of the initial point, following which the dynamics evolve exactly. Alternatively, one can think of robust reachability as a generalisation of the reachability problem, where the target set is required to be reached by a point that is close to the initial point.

5.5.1 The discrete-time robust reachability problem is decidable for non-diagonalisable systems

In this section, we show the full decidability (including for non-diagonalisable systems) of the discrete-time robust reachability problem: decide, given $M \in (\mathbb{R} \cap \overline{\mathbb{Q}})^{L \times L}$, a starting point $x \in \mathbb{Q}^L$ and a target $S \subseteq \mathbb{R}^L$, whether for every $\epsilon > 0$ there exists n and $\delta \in B(0, \epsilon) = \epsilon B(\mathbf{0}, 1)$ such that $M^n(x + \delta) \in S$. As discussed in section 5.2, wlog we can assume that M is in real Jordan form:

$$M = \operatorname{diag}(J_1, \dots J_k, J_{k+1}, \dots, J_d)$$

where for $1 \leq i \leq k$ the block J_i has two non-real eigenvalues and for $k < i \leq d$ the block J_i has one real eigenvalue. We denote the multiplicity and the spectral radius of J_i by σ_i and ρ_i, respectively.

Like section 5.2, the robust reachability problem can be equivalently stated in terms of any full-dimensional set $\mathcal{B}$ that contains $\mathbf{0}$ in its interior (instead of $B(\mathbf{0}, 1)$) as the "control set". That is, for any such set $\mathcal{B}$, the problem of deciding whether $\forall \epsilon. \exists n : (M^n x + \epsilon M^n \mathcal{B}) \cap S \neq \emptyset$ is equivalent to the robust reachability problem. We first give a set $\mathcal{B}$ that is most appropriate for our purposes. Intuitively, the idea is again to eliminate the rotations in M^n so that $M^n \mathcal{B}$ can be defined in a first-order fashion using algebraic parameters. Let $\mathcal{B} = \Pi_{i=1}^d \mathcal{B}_i$ where (i) $\mathcal{B}_i = \Pi_{j=1}^{\sigma_i} B((0,0), 1)$ for $1 \leq i \leq k$ and (ii) $B_i = [-1, 1]^{\sigma_i}$ for $k < i \leq d$. Define $\mathcal{B}(n) = M^n \mathcal{B}$ and observe that

$$\mathcal{B}(n) = \varphi(n, \rho_1^n, \dots, \rho_d^n)$$

where φ is a semialgebraic function. Let $\widetilde{\mathcal{O}}_\epsilon(n) = M^n x + \epsilon \mathcal{B}(n)$. The robust reachability problem is then equivalent to determining whether

$$\forall \epsilon > 0. \exists n : \widetilde{\mathcal{O}}_\epsilon(n) \cap S \neq \emptyset.$$

We move on to defining the abstraction for $M^n x$. Assume M is of the same form as above. For $\alpha \in \mathbb{T}$ let $R(\alpha) = \begin{bmatrix} \operatorname{Re}(\alpha) & -\operatorname{Im}(\alpha) \\ \operatorname{Im}(\alpha) & \operatorname{Re}(\alpha) \end{bmatrix}$ and for $1 \leq i \leq k$ let $\gamma_i = \lambda_i / \rho(J_i)$ for a non-real eigenvalue λ_i of the block J_i. Let $f : \mathbb{N} \times \mathbb{T}^k \to \mathbb{R}^{L \times L}$ be the "matrix builder" function, defined as follows.

$$f(n, (\alpha_1, \dots, \alpha_k)) = \operatorname{diag}(g(\alpha_1), \dots, g(\alpha_k), J_{k+1}^n, \dots, J_d^n), \tag{5.1}$$

where

$$g(\alpha_i) = \begin{bmatrix} A_i & n\Lambda_i^{-1}A_i & \cdots & \binom{n}{\sigma_i-1}\Lambda_i^{-\sigma_i+1}A_i \\ & A_i & \ddots & \vdots \\ & & \ddots & n\Lambda_i^{-1}A_i \\ & & & A_i \end{bmatrix} \text{ and } A_i = \rho_i^n R(\alpha).$$

We define the matrix builder with respect to the state matrix M which is given in real JNF. For example, for $\alpha, \beta \in \mathbb{T}$ and the state matrix

$$M = \begin{bmatrix} \Lambda_1 & I & & & & \\ & \Lambda_1 & & & & \\ & & \Lambda_2 & I & & \\ & & & \Lambda_2 & & \\ & & & & \rho_3 & 1 \\ & & & & & \rho_3 \end{bmatrix},$$

the corresponding matrix builder takes the form

$$f(n, (\alpha, \beta)) = \begin{bmatrix} A & n\Lambda_1^{-1}A & & & & \\ & A & & & & \\ & & B & n\Lambda_2^{-1}B & & \\ & & & B & & \\ & & & & \rho_3^n & n \\ & & & & & \rho_3^n \end{bmatrix}.$$

Here $A = \rho(\Lambda_1)^n R(\alpha)$ and $B = \rho(\Lambda_2)^n R(\beta)$. Let

$$\mathcal{T} = \mathrm{cl}(\{(\gamma_1^n, \ldots, \gamma_k^n) : n \in \mathbb{N}\}).$$

The set $\mathcal{T}$ is semialgebraic and effectively computable. Further define

$$\mathcal{A}_\epsilon(n)(z) = f(n, z)x + \epsilon\mathcal{B}(n) \text{ and } \mathcal{A}_\epsilon(n) = \{\mathcal{A}_\epsilon(n)(z) : z \in \mathcal{T}\}.$$

Then $M^n x = f(n, (\gamma_1^n, \ldots, \gamma_k^n))x$ is abstracted by $\{f(n, z)x : z \in \mathbb{T}^k\}$ and $\mathcal{A}_\epsilon(n) \supseteq \widetilde{\mathcal{O}}_\epsilon(n)$.

The following lemma encapsulates all the nasty differences between the diagonalisable and the non-diagonalisable case. Its proof is an easy manipulation of matrices.

Lemma 5.11. *Given $z = (\alpha_1, \ldots, \alpha_k)$, a time step n, an update matrix M and a starting point x,*

$$f(n, (\alpha_1, \ldots, \alpha_k))x = M^n x + M^n \Delta$$

has a solution

$$\Delta = (\Delta_1, \ldots, \Delta_k, \mathbf{0}, \ldots, \mathbf{0})$$

where $\Delta_i(j) = R(\gamma_i^{-n})(R(\alpha_i) - R(\gamma_i^n))x_i(j)$ for $1 \leq i \leq k$ and $1 \leq j \leq \sigma_i$.

Observe that, assuming k and x are fixed, $||\Delta|| = O(||z - \Gamma^n||)$, where $\Gamma^n = (\gamma_1^n, \ldots, \gamma_k^n)$. Hence we obtain the following corollary, which intuitively states that if z is close to Γ^n, then $f(n,z)x$ is close to the true point $M^n x$, in the sense that $f(n,z)x$ can be reached from x by first jumping to a point x' that is at most $\epsilon\mathcal{B}$ away and then applying M exactly n times.

Corollary 5.12. *Given M and x, there exists C such that for all $\epsilon > 0$ and $z \in \mathbb{T}^k$,*

$$||\Gamma^n - z|| < C\epsilon \implies \exists \Delta \in \epsilon\mathcal{B} : \ f(n,z)x = M^n x + M^n \Delta.$$

We now move onto proving decidability of the robust reachability problems. Observe that the dichotomy lemma (Lemma 5.3) and its proof hold verbatim.

Lemma 5.13. *Either*

1. *for every $\epsilon > 0$ there exists N_ϵ such that for all $n > N_\epsilon$, $\mathcal{A}_\epsilon(n)$ intersects S, or*
2. *there exist computable N and $\epsilon > 0$ such that $\mathcal{A}_\epsilon(n)$ does not intersect S for all $n > N$.*

Moreover, it can be effectively determined which case holds.

If Case 2 holds, then S is robust reachable if and only if it is reachable within the first N steps. We will show that in Case 1, S is robust reachable. This will conclude the proof.

Lemma 5.14. *If for all $\epsilon > 0$ there exists N_ϵ such that for all $n > N_\epsilon$, $\mathcal{A}_\epsilon(n)$ intersects S then S is robust reachable.*

Proof. Let $\epsilon > 0$. We show that S is "ϵ-robust-reachable". That is, $M^n x + \epsilon M^n \mathcal{B}$ intersects S for some n. Consider $\mathcal{A}_{\epsilon/2}$. By assumption, there exists N_1 such that for all $n > N_1$, $\mathcal{A}_{\epsilon/2}(n)$ intersects S. Let ϵ' be sufficiently small such that for all z

$$||\Gamma^n - z|| < \epsilon' \implies \exists \Delta \in \frac{\epsilon}{2}\mathcal{B} : \ f(n,z)x = M^n x + M^n \Delta.$$

Consider the limiting shape L for the sequence

$$T_n = \{z \in \mathcal{T} : \mathcal{A}_{\epsilon/2}(n)(z) \text{ intersects } S\}.$$

By Lemma 5.4 (b), there exists $N > N_1$ such that for all $n > N$, $p + B(\mathbf{0}, \epsilon'/2)$ intersects T_n. That is, for all $n > N$ there exists $p_n \in \mathcal{T}$ such that $||p - p_n|| < \epsilon'/2$ and $p_n \in T_n$. Equivalently,

$$||p - p_n|| < \epsilon'/2 \text{ and } \mathcal{A}_{\epsilon/2}(n)(p_n) \text{ intersects } S.$$

By Kronecker's theorem there must exist $m > N$ such that $||\Gamma^m - p|| < \epsilon'/2$. Hence we have

$$||\Gamma^m - p_m|| \leq \epsilon' \text{ and } \mathcal{A}_{\epsilon/2}(m)(p_m) \text{ intersects } S.$$

By the construction of ϵ' there exists $\Delta \in \frac{\epsilon}{2}\mathcal{B}$ such that $f(n, p_m)x = M^n x + M^n \Delta$. Hence

$$M^n x + \epsilon M^n \mathcal{B} = M^n x + \frac{\epsilon}{2} M^n \mathcal{B} + \frac{\epsilon}{2} M^n \mathcal{B} \supseteq f(n, p_m)x + \frac{\epsilon}{2}\mathcal{B} = \mathcal{A}_{\epsilon/2}(m)(p_m).$$

Since $\mathcal{A}_{\epsilon/2}(m)(p_m)$ intersects S, it follows that $M^n x + \epsilon M^n \mathcal{B}$ intersects S too. □

Theorem 5.15 (Discrete robust reachability)**.** *The discrete robust reachability problem for linear dynamical systems is decidable.*

Proof. We use essentially the same algorithm as described in subsection 5.2.4 with our new abstraction. □

5.5.2 The continuous-time robust reachability problem

In this section, we take the continuous robust reachability problem for linear dynamical systems and show that it can be reduced to the bounded-time reachability problem for semialgebraic target sets. The techniques we use are very similar to the discrete setting. The only major difference is that deciding the continuous-time robust reachability problem requires Schanuel's conjecture.

Let $M \in (\mathbb{R} \cap \overline{\mathbb{Q}})^{L \times L}$ be a matrix in real JNF, $x \in \mathbb{Q}^L$ be a starting point, and $S \subseteq \mathbb{R}^L$ be a semialgebraic target set. We want to show how to decide whether $\forall \epsilon > 0.\, \exists t \in \mathbb{R}_{\geq 0}, \delta \in B(\mathbf{0}, \epsilon) \colon e^{Mt}(x + \delta) \in S$.

Let $\mathcal{B}$ be the control set as defined for the discrete setting. This gives $\mathcal{B}(t) = e^{Mt}\mathcal{B}$ and the ϵ-pseudo-orbit of the continuous system at time $t \in \mathbb{R}_{\geq 0}$ can be defined as $\widetilde{\mathcal{O}}_\epsilon(t) = e^{Mt}x + \mathcal{B}(t)$. The continuous robust reachability problem is equivalent to determining the truth of the sentence

$$\forall \epsilon > 0.\, \exists t \in \mathbb{R}_{\geq 0} \colon \widetilde{\mathcal{O}}_\epsilon(t) \cap S \neq \emptyset.$$

One can define an abstraction for $e^{Mt}x$ similar to the discrete case. In particular, for $\alpha \in \mathbb{R} \cap \overline{\mathbb{Q}}$ let $R(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$ and for $1 \leq i \leq k$ let $\gamma_i = \operatorname{Im}(\lambda_i)$ and $r_i = \operatorname{Re}(\lambda_i)$ for a non-real eigenvalue λ_i of the block J_i. Finally, let $f : \mathbb{R}_{\geq 0} \times \mathbb{R}^k \to \mathbb{R}^{L \times L}$ be the "matrix builder" function, defined as follows.

$$f(t, (\alpha_1, \ldots, \alpha_k)) = \operatorname{diag}(g(\alpha_1), \ldots, g(\alpha_k), e^{J_{k+1}t}, \ldots, e^{J_d t})$$

where

$$g(\alpha_i) = \begin{bmatrix} A_i & tA_i & \cdots & \frac{t^{\sigma_i - 1}}{(\sigma_i-1)!}A_i \\ & A_i & \ddots & \vdots \\ & & \ddots & tA_i \\ & & & A_i \end{bmatrix} \text{ and } A_i = e^{r_i t} R(\alpha).$$

Let

$$\mathcal{T} = \{(R(\gamma_1 t), \ldots, R(\gamma_k t)) : t \in \mathbb{R}_{\geq 0}\}.$$

The set $\mathcal{T}$ is semialgebraic and effectively computable. Further define

$$\mathcal{A}_\epsilon(t)(z) = f(t, z)x + \epsilon\mathcal{B}(t) \text{ and } \mathcal{A}_\epsilon(t) = \{\mathcal{A}_\epsilon(t)(z) : z \in \mathcal{T}\}.$$

Notice that we have $\mathcal{A}_\epsilon(n) \supseteq \widetilde{\mathcal{O}}_\epsilon(n)$. The next lemma shows that if z is picked close enough to $\Gamma(t) = (\gamma_1 t, \ldots, \gamma_k t)$, then $e^{Mt}x$ can be approximated by $f(t, z)x$.

Lemma 5.16. *Given $z = (\alpha_1, \ldots, \alpha_k)$, a time point $t \in \mathbb{R}_{\geq 0}$, an update matrix M and a starting point x,*

$$f(t, (\alpha_1, \ldots, \alpha_k))x = e^{Mt}x + e^{Mt}\Delta$$

has a solution

$$\Delta = (\Delta_1, \ldots, \Delta_k, \mathbf{0}, \ldots, \mathbf{0})$$

where $\Delta_i(j) = R(-\gamma_i t)(R(\alpha_i) - R(\gamma_i t))x_i(j)$ for $1 \leq i \leq k$ and $1 \leq j \leq \sigma_i$.

We have the following corollary, similar to the discrete setting.

Corollary 5.17. *Given M and x, there exists C such that for all $\epsilon > 0$ and $z \in (\mathbb{R} \cap \overline{\mathbb{Q}})^k$,*

$$||\Gamma(t) - z|| < C\epsilon \implies \exists \Delta \in \epsilon\mathcal{B} : \; f(t, z)x = e^{Mt}x + e^{Mt}\Delta.$$

Theorem 5.18. *The continuous robust reachability problem for linear dynamics reduces to bounded-time reachability problem for linear continuous-time systems.*

Proof. The proof is a direct combination of the proofs for continuous pseudo-reachability (section 5.4) and discrete robust reachability (subsection 5.5.1). □

As before, since the bounded-time reachability problem for continuous-time linear dynamical systems can be encoded in $\mathbb{R}_{\exp,\cos\restriction[0,T]}$ (or solved directly assuming Schanuel [21]), we have the following (conditional) decidability result.

Corollary 5.19. *The continuous robust reachability problem for the continuous-time linear dynamical systems is decidable subject to Schanuel's conjecture.*

5.6 Conclusion

The main technical result of this chapter is that it is decidable whether

$$\forall \epsilon > 0.\, \exists n : (M^n x + f(n) + \epsilon \mathcal{B}(n)) \cap S \neq 0$$

where M is a diagonalisable matrix with algebraic entries, x is an algebraic starting point, f is a semialgebraic function, S is a semialgebraic target and $\mathcal{B}(n) = \varphi(n, \rho_1^n, \ldots, \rho_d^n)$ for $\rho_1, \ldots, \rho_d \in \mathbb{R} \cap \overline{\mathbb{Q}}$ and a semialgebraic function φ.

We used this result to show decidability of the discrete-time pseudo-reachability problem for diagonalisable systems in the following way. We first observed that the pseudo-reachability problem can be cast as the problem of determining whether $\forall \epsilon > 0.\, \exists n : \widetilde{\mathcal{O}}_\epsilon(n) \cap S \neq \emptyset$, where $\widetilde{\mathcal{O}}_\epsilon(n)$ is the set of all points that are reachable exactly at the time n via an ϵ-pseudo-orbit. After choosing $\mathcal{B}$ as the most convenient control set (see Lemma 5.2 and subsection 5.2.1), we then showed that $\widetilde{\mathcal{O}}_\epsilon(n)$ can be written as $M^n x + f(n) + \epsilon \mathcal{B}(n)$.

The reason we are unable to show decidability for non-diagonalisable systems in this fashion is that we are unable to write $\widetilde{\mathcal{O}}_\epsilon(n)$ as $M^n x + f(n) + \epsilon \mathcal{B}(n)$. For example, already for the Jordan block $M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, and in general already for blocks with a single repeated real eigenvalue, we do not know whether it is even possible to eliminate the summation $\sum_{i=0}^{n-1} M^i B$ and express $\widetilde{\mathcal{O}}_\epsilon(n) = M^n x + \epsilon \sum_{i=0}^{n-1} M^i B$, where B is any full dimensional shape containing $\mathbf{0}$ in its interior, in the required fashion.

Our approach, however, can be used to solve, in full generality, the robust reachability problem of [1]: given M, x and S, decide whether $\forall \epsilon > 0 : \exists n : (M^n x + \epsilon M^n B(\mathbf{0}, 1)) \cap S \neq \emptyset$. Intuitively, the reason is that in this version there is no summation of the form $\sum_{i=0}^{n-1} M^i B$. For diagonalisable systems in particular, decidability of the robust reachability problem is almost immediate. First, one can again show that the problem is equivalent to determining whether $\forall \epsilon > 0 : \exists n : (M^n x + \epsilon M^n \mathcal{B}) \cap S \neq \emptyset$. It then remains to observe that $M^n \mathcal{B} = \varphi(n, \rho_1^n, \ldots, \rho_d^n)$ for a semialgebraic predicate φ and apply the technical result described above.

The same techniques also apply directly to the continuous analogues of these problems, with the intriguing difference that reducing to bounded reachability is a theoretically trivial finite case check in the discrete case but not in the continuous case. In the continuous case we run into the classic problem of root-finding zeroes of exponential-polynomial functions. The issue is that a zero may be tangential, thus similarly to the difficulty in deciding zeroness of an elementary number, it is not clear how to confirm the existence of such a zero via finite-precision numerical computation.

The paper [21] gives a method to factorise an exponential polynomial into factors that do not have tangential zeroes assuming Schanuel's conjecture. However this does not give a bound on the precision required to find the non-tangential zeroes (that would require a quantitative version of Schanuel's conjecture) so we do not have a complexity bound for this problem.

Chapter 6

Bounding the Escape Time of a Linear Dynamical System over a Compact Semialgebraic Set

6.1 Introduction

This final chapter has a different flavour to the rest of this thesis. While previously we have considered decidability of reachability properties, which correspond to safety, we switch gears and investigate liveness. We also go beyond decidability and consider quantification of the time it takes for a system to escape a given set.

An *invariant set* of a dynamical system is a set K such that every trajectory that starts in K remains in K. Dually, an *escape set* K is one such that every trajectory that starts in K eventually leaves K (either temporarily or permanently). While it is usually straightforward to establish that a given set K is invariant, it can be challenging to decide whether K is an escape set. Indeed, while the former problem amounts to showing that K is closed under the transition function, the latter potentially involves considering entire orbits. In particular, even in case K has a finite escape time (the maximum number of steps for an orbit to escape the set), it can be highly non-trivial to establish an explicit upper bound on the escape time.

In this chapter we focus on escape sets for (discrete-time) linear dynamical systems. Given a rational matrix $A \in \mathbb{Q}^{n\times n}$ we say that $K \subseteq \mathbb{R}^n$ is an escape set for A if for all points $x \in K$, there exists $t \in \mathbb{N}$ such that $A^t x \notin K$. The *compact escape problem (CEP)* asks to decide whether a given compact semialgebraic set K is an escape set for a given matrix A. Decidability of CEP was shown in [71] and its computational complexity was characterised in [30] as being interreducible with the decision problem for a certain fragment of the theory of real closed fields.

Here we focus exclusively on positive instances (A, K) of CEP, that is, we assume that we are given a compact semialgebraic escape set for a linear dynamical system. In this situation it turns out, due to compactness of K, that there exists a finite time T such that for all $x \in K$ there exists $t \leq T$ with $A^t x \notin K$. The least such T is called the *escape time* of (A, K). Our main result (Theorem 6.1, shown below) gives an explicit upper bound on the escape time of (A, K) as a function of the length of the description of the matrix A and semialgebraic set K. In general, it is recognised that bounded liveness is a more useful property than mere liveness. Theorem 6.1 can be used to establish bounded liveness of several kinds of systems. For example, the result gives an upper bound on the termination time of a single-path linear loop with compact guard (cf. [92, 17]); it also gives a bound on the number of steps to remain in a particular control location of a hybrid system before a given (compact) state invariant becomes false, forcing a transition.

We next recall some terminology to formalise our main contribution. We say that a semialgebraic set S has *description size* at most (n, d, τ) if it can be expressed by a boolean combination of polynomial equations and inequalities $P(x_1, \ldots, x_n) \bowtie 0$ with $\bowtie \in \{\leq, =\}$, involving polynomials $P \in \mathbb{Z}[x_1, \ldots, x_n]$ in at most n variables of total degree at most d with integer coefficients bounded in bitsize by τ. Our main result is as follows:

Theorem 6.1. *There exists a function* $\mathrm{CompactEscape}(n, d, \tau) \in 2^{(d\tau)^{n^{O(1)}}}$ *with the following property. If $K \subseteq \mathbb{R}^n$ is a compact semialgebraic set of description size at most (n, d, τ) that is an escape set for a matrix $A \in \mathbb{Q}^{n \times n}$ with entries of bitsize at most τ, then the escape time of K is bounded by* $\mathrm{CompactEscape}(n, d, \tau)$.

As explained in the proof sketch below, Theorem 6.1 relies on the availability of certain quantitative bounds within semialgebraic geometry and number theory, particularly concerning quantifier elimination and Diophantine approximation. The latter results are crucial to handling the case in which the matrix A has complex eigenvalues of absolute value one.

Note that the upper bound on the escape time in Theorem 6.1 is singly exponential in the degrees and the bitsize of the coefficients of the polynomials used to define K and the bitsize of the coefficients of A. It is doubly exponential in the dimension. In Section 6.6 we provide two examples that yield a corresponding lower bound of this form. It is moreover straightforward to give examples of non-compact escape sets for which there is no finite uniform escape time bound, i.e., for all $T \in \mathbb{N}$ there exists $x \in K$ that still escapes but requires greater than T steps.

6.1.1 Proof Overview

Let us now give a high-level overview of the proof of Theorem 6.1. As in the statement of the theorem, let $K \subseteq \mathbb{R}^n$ be a compact semialgebraic set of description size at most (n, d, τ) and let $A \in \mathbb{Q}^{n\times n}$ be a matrix with entries of bitsize bounded by τ, and such that for all $x \in K$ there exists $t \in \mathbb{N}$ such that $A^t x \notin K$.

To facilitate the analysis of the dynamical behaviour of A we first transform our system into real Jordan normal form using Lemma 2.19. A theorem of Cai [18] ensures that this step does not significantly increase the description size of the system.

The dynamics of A naturally decomposes into a rotational part, corresponding to eigenvalues of modulus one, and an expansive or contractive part, corresponding to eigenvalues of absolute value different from 1 and to generalised eigenvalues of arbitrary moduli. Accordingly, the ambient space $\mathbb{R}^n$ decomposes into two subspaces V_{rec} and $V_{\text{non-rec}}$, such that A exhibits rotational behaviour on V_{rec} and expansive or contractive behaviour on $V_{\text{non-rec}}$. We start by considering the special cases where either $V_{\text{rec}} = 0$ or $V_{\text{non-rec}} = 0$, so that only one of the two types of behaviours occurs.

First, assume that A has no complex eigenvalues of modulus 1. Since every trajectory under A escapes K we have in particular that $0 \notin K$. A theorem due to Jeronimo, Perrucci and Tsigaridas [52] shows that K is bounded away from zero by a function of the form $2^{-(d\tau)^{n^{O(1)}}}$ and a theorem due to Vorobjov [96] establishes an upper bound on the absolute value of every coordinate of every point in K of the form $2^{(d\tau)^{n^{O(1)}}}$. Furthermore, thanks to a result of Mignotte [68], we can bound the eigenvalues of A away from 1 by a function of the form $2^{\tau^{n^{O(1)}}}$. This yields a doubly exponential bound on how long it takes for A to leave the set K (either by converging to 0 or by converging to infinity in some eigenspace).

Now assume that all eigenvalues of A have modulus 1. This case is handled through a combination of two bounds. For the first bound we start by noting that for every $x \in K$ the closure of the orbit $\overline{\mathcal{O}_A(x)}$ is a compact semialgebraic set that is not entirely contained within K. In fact we show that for all $x \in K$ there exists a point $y \in \overline{\mathcal{O}_A(x)}$ whose distance to K is at least $2^{-(d\tau)^{n^{O(1)}}}$. This bound is achieved by applying [52, Theorem 1] to a suitable polynomial on an auxiliary semialgebraic set, which is constructed using quantifier elimination. The singly exponential bounds obtained in [50, 82] are crucial for this step to work. The second step of the argument applies a quantitative version of Kronecker's theorem on simultaneous Diophantine approximation to obtain a bound of the form $N_P \in 2^{(\tau P)^{n^{O(1)}}}$ such that for all positive integers P every point $z \in \overline{\mathcal{O}_A(x)}$ is within 2^{-P} of a point of the form $A^t x$ with

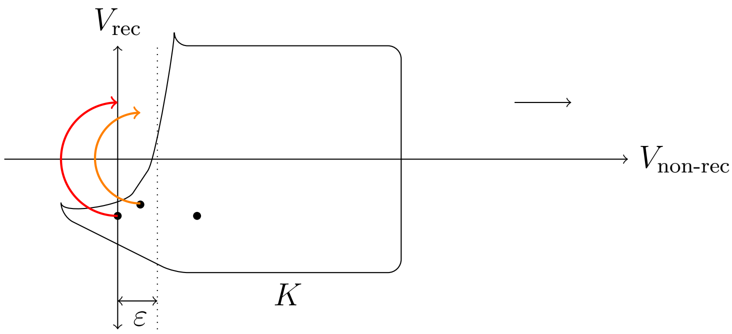

Figure 6.1: The red arrow represents a point which escapes rotationally while the blue arrow represents a point which escapes expansively. The orange arrow represents a point which escapes rotationally due to being close enough to V_{rec} that it shares the behavior of points in V_{rec}.

$0 \leq t \leq N_P$. Combining the two bounds described above, we obtain a doubly exponential bound on the escape time.

In the presence of both types of behaviour, the analysis of each case becomes more involved. We select a parameter $\varepsilon > 0$ and partition K into three sets: $K_{\text{rec}} = K \cap V_{\text{rec}}$, $K_{\geq\varepsilon}$, and $K_{<\varepsilon}$. The matrix A exhibits purely rotational behaviour on K_{rec}. Intuitively, on $K_{\geq\varepsilon}$ the expansive or contractive behaviour of A dominates the overall dynamics, while on $K_{<\varepsilon}$ the rotational behaviour dominates the overall dynamics. Such behavior is illustrated in Figure 6.1. We establish in Proposition 6.2 a bound N_{rec} such that for each initial point $x \in V_{\text{rec}}$, one of its first N_{rec} iterates is bounded away from K. In Proposition 6.8 we establish a bound $N_{\geq\varepsilon}$ such that every $x \in K_{\geq\varepsilon}$ either escapes or enters $K_{<\varepsilon} \cup K_{\text{rec}}$ within at most $N_{\geq\varepsilon}$ iterations. Finally, in Section 6.5, we establish a bound on how often the system can switch from a state where rotational behaviour dominates to one where expansive or non-expansive behaviour does and vice versa. We use this to combine the two bounds to an overall bound on the escape time, proving Theorem 6.1.

6.2 Decomposing K in the Jordan norm

To obtain a bound on the escape time it will be important to work with instances of the Escape Problem in real Jordan normal form. This transformation does not greatly increase the description size of the system thanks to Lemma 2.19. Once in Jordan form, we shall partition K using a Jordan norm defined below. The idea is similar to the "rotation-invariance" in subsection 5.2.1 - we want a norm in which the matrix is an isometry on a subspace of $\mathbb{R}^n$.

Let $K \subseteq \mathbb{R}^n$ be a compact semialgebraic set. Let $A \in \mathbb{R}^{n\times n}$ be a matrix in real Jordan normal form,

$$A = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix}.$$

Here, each J_i is a real Jordan block of the form

$$J_i = \begin{pmatrix} \Lambda_i & I_i & & \\ & \ddots & \ddots & \\ & & \Lambda_i & I_i \\ & & & \Lambda_i \end{pmatrix},$$

where $\Lambda_{i,1}$ is either a real number or a 2×2 real matrix of the form $\begin{pmatrix} a_i & -b_i \\ b_i & a_i \end{pmatrix}$ and, accordingly, I_i is either the real number 1 or the 2×2 identity matrix. The elements Λ_i correspond to real or complex eigenvalues $\lambda_i \in \mathbb{C}$ of A. By slight abuse of language we call $|\lambda_i|$ the modulus of Λ_i. By further slight abuse of language we define the "eigenspace" of Λ_i as the one- or two-dimensional space spanned by the vectors that correspond to the first entry of the Jordan block J_i. The "generalised eigenspaces" for Λ_i are defined analogously.

Write $\mathbb{R}^n$ as the direct sum of two spaces $\mathbb{R}^n = V_{\text{rec}} \oplus V_{\text{non-rec}}$ where V_{rec} is the direct sum of the eigenspaces for eigenvalues of modulus 1, and $V_{\text{non-rec}}$ is the direct sum of the eigenspaces and generalised eigenspaces for eigenvalues of modulus $\neq 1$ and the generalised eigenspaces for eigenvalues of modulus 1. By convention, if A has no eigenvalues of modulus 1 we let $V_{\text{rec}} = 0$. Similarly, if A has only eigenvalues of modulus 1 and no generalised eigenvalues we let $V_{\text{non-rec}} = 0$. Thus, we decompose the state space $\mathbb{R}^n$ into a part V_{rec} on which A exhibits purely rotational behaviour, and a part $V_{\text{non-rec}}$ where A is additionally expansive or contractive.

In addition to the familiar ℓ^2 and ℓ^∞ norms we now introduce a third norm, depending on the matrix A, that combines features of the two. It facilitates block-wise arguments while ensuring that the restriction of A to V_{rec} is an isometry.

Write $\mathbb{R}^n$ as a direct sum $\mathbb{R}^n = V_1 \oplus \cdots \oplus V_s \oplus W_1 \oplus \cdots \oplus W_t$, where $V_1, \ldots, V_s$ correspond to the Jordan blocks of A associated with real eigenvalues and $W_1, \ldots, W_t$ correspond to the Jordan blocks of A associated with non-real eigenvalues. Let $\pi_{W_j} \colon \mathbb{R}^n \to W_j$ and $\pi_{V_j} \colon \mathbb{R}^n \to V_j$ denote the orthogonal projections onto W_j and V_j respectively.

For a vector $x \in V_i$, let $\|x\|_J^{V_i} = \|x\|_\infty$. For a vector $x = (x_1, y_1, \ldots, x_k, y_k) \in W_i$, let

$$\|x\|_J^{W_i} = \max_{j=1,\ldots,k} \left(\sqrt{x_j^2 + y_j^2}\right).$$

For a vector $x \in \mathbb{R}^n$, let

$$\|x\|_J = \max\left\{ \max_{j=1,\dots,s} \left\|\pi_{V_j}(x)\right\|_J^{V_j}, \max_{j=1,\dots,t} \left\|\pi_{W_j}(x)\right\|_J^{W_j} \right\}.$$

Call $\|x\|_J$ the Jordan norm of x. Observe that $\|x\|_J$ depends on the choice of the V_i's and W_i's. The Jordan norm compares to the ℓ^2- and ℓ^∞- norms as follows:

$$n^{-1/2}\|x\|_J \le n^{-1/2}\|x\|_2 \le \|x\|_\infty \le \|x\|_J \le \|x\|_2 \le n^{1/2}\|x\|_\infty \le n^{1/2}\|x\|_J.$$

Let $\varepsilon > 0$. Consider the ball $B_J(0,\varepsilon) \subseteq \mathbb{R}^n$ about 0 with respect to the distance induced by the $\|\cdot\|_J$-norm. We partition K into three sets:

$$\begin{aligned} K_{\text{rec}} &:= K \cap V_{\text{rec}} \\ K_{<\varepsilon} &:= K \cap (V_{\text{rec}} \oplus ((V_{\text{non-rec}} \cap B_J(0,\varepsilon)) \setminus \{0\})) \\ K_{\ge\varepsilon} &:= K \cap (V_{\text{rec}} \oplus (V_{\text{non-rec}} \setminus B_J(0,\varepsilon))). \end{aligned}$$

6.3 The recurrent eigenspace bound

The next proposition establishes as a special case an escape bound for all initial values $x \in K_{\text{rec}}$. In order to combine the recurrent and the non-recurrent case we need a stronger result, however. Thus, we establish not only a bound on the escape time for all initial values $x \in K_{\text{rec}}$, but a bound N such that every $x \in V_{\text{rec}}$ (not just in K_{rec}) has a point in its orbit with distance at least $1/N$ – not just positive distance – from K. Further, note that Proposition 6.2 is still applicable in the special cases where $K_{\text{rec}} = \emptyset$ or $V_{\text{rec}} = 0$.

Proposition 6.2 (recurrent escape bound). *There exists a function* $\mathrm{Rec}(n,d,\tau) \in 2^{(\tau d)^{n^{O(1)}}}$ *with the following property:*

Let $A \in \mathbb{A}^{n\times n}$ be a matrix in real Jordan normal form with algebraic entries. Assume that the minimal polynomial of A has rational coefficients whose bitsize is bounded by τ. Let $K \subseteq \mathbb{R}^n$ be a compact semialgebraic set of description size at most (n,d,τ). If every point $x \in K_{\text{rec}}$ escapes K under iterations of A then for all $x \in V_{\text{rec}}$ there exists $t \le \mathrm{Rec}(n,d,\tau)$ such that

$$\mathrm{dist}_{\ell^2}(A^t x, K) > \frac{\sqrt{n}}{\mathrm{Rec}(n,d,\tau)}.$$

Proof. We sketch an outline here. The rest of this section is devoted to fleshing out this outline. Note that we introduce the unexpected factor of $\sqrt{n}$ in the numerator in order to later facilitate conversions between the Jordan norm and the standard ℓ^2 norm.

We first prove the result for initial points $x \in K_{\mathrm{rec}}$. For these points, the closure of the orbit $\overline{\mathcal{O}_A(x)}$ of x under A is a compact semialgebraic set. We employ Theorem 2.18 to obtain for all $\varepsilon > 0$ a doubly exponential bound N such that for all $x \in K_{\mathrm{rec}}$ and all $y \in \overline{\mathcal{O}_A(x)}$ there exists $t \leq N$ such that $\|A^t x - y\|_2 < \varepsilon$. We then use Theorem 2.14 to obtain a uniform at most doubly exponentially small lower bound on the quantity

$$\inf_{x \in K_{\mathrm{rec}}} \sup_{y \in \overline{\mathcal{O}_A(x)}} \inf_{z \in K} \|y - z\|_2^2 .$$

In order to apply this theorem we construct an auxiliary semialgebraic set, whose description size is controlled by Theorem 2.12. Combining these two steps, we obtain a function Rec_0 that satisfies the statement of the lemma for all initial points $x \in K_{\mathrm{rec}}$.

Finally, we extend the result to all initial points $x \in V_{\mathrm{rec}}$ as follows:

The special case where $K_{\mathrm{rec}} = \emptyset$ is treated using Theorem 2.14. In the case where K_{rec} is non-empty we obtain from Lemma 6.7 that a point in V_{rec} can't stay close to K while being far from K_{rec}. More specifically, every $x \in V_{\mathrm{rec}}$ which is doubly exponentially close to K with a sufficiently large constant in the third exponent is already doubly exponentially close to K_{rec}, with a slightly smaller constant in the third exponent.

Now, any point in V_{rec} that is sufficiently far away from K trivially satisfies our proposition. By the preceding discussion, points $x \in V_{\mathrm{rec}}$ that are sufficiently close to K are already sufficiently close to K_{rec}, so that there exists an escaping orbit $\overline{\mathcal{O}_A(x')}$ with $x' \in K_{\mathrm{rec}}$ which is close to the orbit of x since A is an isometry on V_{rec}. This allows us to reduce the result to the already established result for initial values in K_{rec}. □

6.3.1 Initial points in K_{rec}

We will first prove the following weaker version of Proposition 6.2, where we only establish an escape bound and a lower bound on the distance to K for initial points in K_{rec}.

Lemma 6.3. *There exists a function* $\mathrm{Rec}_0(n, d, \tau) \in 2^{(\tau d)^{n^{O(1)}}}$ *with the following property:*

Let $A \in \mathbb{A}^{n \times n}$ *be a matrix in real Jordan normal form. Assume that the minimal polynomial of* A *has rational coefficients whose bitsize is bounded by* τ. *Let* $K \subseteq \mathbb{R}^n$ *be a semialgebraic set of description size at most* (n, d, τ). *If every point* $x \in K_{\mathrm{rec}}$

escapes K under iterations of A then for all $x \in K_{\mathrm{rec}}$ there exists $t \le \mathrm{Rec}_0(n,d,\tau)$ such that

$$\mathrm{dist}_{\ell^2}(A^t x, K) > \frac{\sqrt{n}}{\mathrm{Rec}_0(n,d,\tau)}.$$

For $x \in V_{\mathrm{rec}}$, let

$$\mathcal{O}_A(x) = \left\{A^t x \mid t \in \mathbb{N}\right\}$$

denote the orbit of x under A. Let $\overline{\mathcal{O}_A(x)}$ denote its closure. By Kronecker's theorem (Theorem 2.16), the set $\overline{\mathcal{O}_A(x)}$ is semialgebraic.

The sequence $(A^t x)_{t\in\mathbb{N}}$ is dense in $\overline{\mathcal{O}_A(x)}$ by definition. A combination of Theorem 2.18 and Theorem 2.13 yields a quantitative refinement of this qualitative statement:

Lemma 6.4. *There exists a function $\mathrm{D}(n,d,\tau,P) \in 2^{(\tau P d)^{n^{O(1)}}}$ with the following property:*

Let A be a matrix in real Jordan normal form. Assume that the characteristic polynomial of A has rational coefficients whose bitsize is bounded by τ. Let K be a compact semialgebraic set of description size at most (n,d,τ). Let P be a positive integer. Then for all $x \in K_{\mathrm{rec}}$ and all $y \in \overline{\mathcal{O}_A(x)}$ there exists $t \le \mathrm{D}(n,d,\tau,P)$ such that

$$\left\|A^t x - y\right\|_2 < 2^{-P}.$$

Proof. Let $\mathrm{Kron}(n,\tau,P) \in 2^{(\tau P)^{n^{O(1)}}}$ be the function from Theorem 2.18.

Let $\mathrm{Bound}(n,d,\tau) = 2^{\tau d^{\beta(n+1)}}$ be the radius of a ball enclosing K, where β is the constant from Theorem 2.13.

Put

$$\mathrm{D}(n,d,\tau,P) = \mathrm{Kron}\left(n,\tau,P + \lceil \log(n)/2 + \log\left(\mathrm{Bound}(n,d,\tau)\right)\rceil\right).$$

It is easy to see that $\mathrm{D}(n,d,\tau,P) \in 2^{(\tau P d)^{n^{O(1)}}}$ as claimed.

To prove that D has the desired properties, let A and K be a matrix and a compact semialgebraic set as above. Let P be a positive integer. For a matrix $Q = (q_{i,j})_{i,j=1}^n \in \mathbb{R}^{n\times n}$, let

$$\|Q\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n q_{i,j}^2\right)^{1/2}$$

denote the Frobenius norm of Q. The Frobenius norm is sub-multiplicative and hence satisfies

$$\|Q\cdot x\|_2 = \|Q\cdot x\|_F \le \|Q\|_F \cdot \|x\|_F = \|Q\|_F \cdot \|x\|_2$$

for all $x \in \mathbb{R}^n$.

Let $C = \mathrm{Bound}(n, d, \tau)$. Let $x \in K_{\mathrm{rec}}$. Let $y \in \overline{\mathcal{O}_A(x)}$. Then, by Theorem 2.18, there exist $t \leq \mathrm{Kron}(n, \tau, P + \lceil \log(n)/2 + \log C \rceil)$ and $Q \in \mathbb{R}^{n \times n}$ such that $y = Qx$ and

$$\left\|A^t - Q\right\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} (A^t_{i,j} - Q_{i,j})^2\right)^{1/2} < \left(n \cdot 2^{-2(P + \lceil \log(n)/2 + \log C \rceil)}\right)^{1/2} \leq \frac{2^{-P}}{C}.$$

The first inequality holds because $x \in K_{\mathrm{rec}}$ and $y \in \overline{\mathcal{O}_A(x)}$ and thus only the modulus 1 eigenspaces are relevant, so A^t (which is in Jordan form) and Q will actually differ in at most n entries. Hence:

$$\left\|A^t x - y\right\|_2 = \left\|(A^t - Q)x\right\|_2 \leq \left\|A^t - Q\right\|_F \cdot |x| < 2^{-P}.$$

The last inequality uses that by Theorem 2.13 we have $|x| \leq C$. □

By definition, a point $x \in K_{\mathrm{rec}}$ escapes K under iterations of A if and only if there exists $y \in \overline{\mathcal{O}_A(x)}$ such that $\mathrm{dist}_{\ell^2}(y, K) > 0$. Our next goal is to sharpen this to a uniform lower bound on $\inf_{x \in K} \sup_{y \in \overline{\mathcal{O}_A(x)}} \mathrm{dist}_{\ell^2}(y, K)$. The main idea is to employ Theorem 2.14. Since the function $g(x) = \sup_{y \in \overline{\mathcal{O}_A(x)}} \mathrm{dist}_{\ell^2}(y, K)$ is not a polynomial, we construct an auxiliary compact semialgebraic set in higher dimension that allows us to reduce the problem of finding a uniform lower bound on $g(x)$ to the problem of finding a uniform lower bound on a polynomial. The idea is essentially the same as that of the proof of Lemma 2.15.

Let

$$\overline{\mathcal{O}_A} = \left\{(x, y) \in \mathbb{R}^{2n} \;\mid\; x \in K_{\mathrm{rec}}, y \in \overline{\mathcal{O}_A(x)}\right\}.$$

Let

$$S = \left\{(x, y, z) \in \mathbb{R}^{3n} \;\mid\; x \in K_{\mathrm{rec}}, y \in \overline{\mathcal{O}_A(x)}, z \in K\right\} = \overline{\mathcal{O}_A} \times K.$$

By compactness of K, the number

$$\eta = \min_{x \in K_{\mathrm{rec}}} \max_{y \in \overline{\mathcal{O}_A(x)}} \min_{z \in K} \left\|y - z\right\|_2^2 \tag{6.1}$$

is strictly positive. Letting $\mathrm{D}(n, d, \tau, P)$ denote the function from Lemma 6.4, observe that every point $x \in K_{\mathrm{rec}}$ escapes K under at most

$$\mathrm{D}\left(n, d, \tau, \lceil \log\left(1/\eta\right) \rceil\right)$$

iterations of A. To obtain a bound on the escape time it hence suffices to obtain a bound of (6.1) away from zero. This is achieved by expressing (6.1) as the minimum of a polynomial over a compact semialgebraic set. To this end, we consider the set

$$\begin{aligned} S' := \{&(x, y, z) \in \mathbb{R}^{3n} \mid \\ &x \in K_{\mathrm{rec}}, y \in \overline{\mathcal{O}_A(x)}, z \in K, \forall w \in \overline{\mathcal{O}_A(x)}.\, \mathrm{dist}_{\ell^2}(y, K) \geq \mathrm{dist}_{\ell^2}(w, K)\}. \end{aligned}$$

Observe that

$$\min_{x \in K_{\mathrm{rec}}} \max_{y \in \overline{\mathcal{O}_A(x)}} \min_{z \in K} \|y - z\|_2^2 = \min_{(x,y,z) \in S'} \|y - z\|_2^2 . \tag{6.2}$$

This leads to two problems:

1. Bound the description size of $\overline{\mathcal{O}_A}$ (and thus of S) in terms of the description size of A and K.
2. Bound the description size of S' in terms of the description size of S.

6.3.1.1 Description size of $\overline{\mathcal{O}_A}$

Let us first bound the description size of $\overline{\mathcal{O}_A}$ in terms of that of A and K.

Up to suitably permuting the dimensions, which does not affect the description size, we may assume that A takes the following form:

$$A = \begin{pmatrix} I_{m_1 \times m_1} & & & & & \\ & -I_{m_2 \times m_2} & & & & \\ & & \Lambda_1 & & & \\ & & & \ddots & & \\ & & & & \Lambda_{m_3} & \\ & & & & & B \end{pmatrix} \tag{6.3}$$

where $I_{m \times m}$ is the $m \times m$ identity matrix, $\Lambda_1, \ldots, \Lambda_{m_3}$ are 2×2 matrices corresponding to the genuinely complex eigenvalues of A of modulus 1, and B is some matrix. With this convention, K_{rec} is the set of all $x \in K$ with $x_j = 0$ for $j > m_1 + m_2 + m_3$.

Let

$$A_0 = \begin{pmatrix} I_{m_1 \times m_1} & & & & & \\ & -I_{m_2 \times m_2} & & & & \\ & & \Lambda_1 & & & \\ & & & \ddots & & \\ & & & & \Lambda_{m_3} & \\ & & & & & \mathbf{0} \end{pmatrix}$$

where $\mathbf{0}$ is the zero matrix of the same dimensions as B in (6.3). Clearly, the orbits $\mathcal{O}_A(x)$ and $\mathcal{O}_{A_0}(x)$ coincide for all $x \in K_{\mathrm{rec}}$.

Let

$$L = \left\{(\alpha_1, \ldots, \alpha_{m_3}) \in \mathbb{Z}^{m_3} \;\middle|\; \Lambda_1^{\alpha_1} \cdot \cdots \cdot \Lambda_{m_3}^{\alpha_{m_3}} = I_{2\times 2}\right\}.$$

By Theorem 2.5 the abelian group L has a basis $\beta_1, \ldots, \beta_\ell$ with the magnitudes of $\beta_1, \ldots, \beta_\ell$ bounded polynomially in the data that is used to describe the algebraic entries of the matrices $\Lambda_1, \ldots, \Lambda_{m_3}$ and singly exponentially in n. Thus, by our assumption on the characteristic polynomial of A (coefficient bitsize bounded by τ), we have $\sum_{i,j} |\beta_{i,j}| = (n\tau)^{n^{O(1)}}$.

By Kronecker's theorem (Theorem 2.16), the closure of the set $(A_0^t)_{t\in\mathbb{N}}$ in $\mathbb{R}^{n\times n}$ is given by the set of all matrices of the form

$$\begin{pmatrix} I_{m_1\times m_1} & & & & & \\ & \sigma \cdot I_{m_2\times m_2} & & & & \\ & & Z_1 & & & \\ & & & \ddots & & \\ & & & & Z_{m_3} & \\ & & & & & \mathbf{0} \end{pmatrix},$$

where $\sigma \in \{-1, 1\}$ and $Z_1, \ldots, Z_{m_3}$ are 2×2 matrices satisfying

$$Z_1^{\beta_{i,1}} \cdot \cdots \cdot Z_{m_3}^{\beta_{i,m_3}} = I_{2\times 2} \tag{6.4}$$

for all $i = 1, \ldots, \ell$.

The size of the polynomials required to describe the relation (6.4) can be bounded by the following straightforward lemma:

Lemma 6.5. *Let $M_1, \ldots, M_N$ be 2×2 matrices with entries in a polynomial ring $\mathbb{Z}[x_1, \ldots, x_k]$. Let d be a bound on the total degree of all matrix entries. Let τ be a bound on the bitsize of the coefficients of all matrix entries. Then the entries of the product $M_1 \cdot \cdots \cdot M_N$ have total degree at most Nd and coefficients bounded in bitsize by $N\tau + k(N-1)\log(d+1)$.*

Proof. By induction on N. For $N = 1$ there is nothing to prove. Assume the claim holds true for $N \in \mathbb{N}$. Consider the product $M_1 \cdot \cdots \cdot M_N \cdot M_{N+1}$.

Write $M = M_1 \cdot \cdots \cdot M_N$. By induction hypothesis the entries of the matrix M have total degree at most Nd and coefficients bounded in bitsize by $N\tau + k(N-1)\log(d+1)$.

All entries of $M \cdot M_{N+1}$ are of the form $PQ + RS$ with P and R entries of M and Q and S entries of M_{N+1}. Using standard estimates on the bitsize of sum and product of integers we obtain that this polynomial has degree at most $Nd+d=(N+1)d$ and coefficients bounded in bitsize by $(N+1)\tau+kN\log(d+1)$. $\square$

Thus, the closure $\mathcal{M}$ of $(A_0^t)_{t\in\mathbb{N}}$ is an algebraic set that can be described using at most n^2 polynomials in n^2 variables, each of which has its degree bounded by $(n\tau)^{n^{O(1)}}$, and coefficients bounded in bitsize by a polynomial in n and τ.

We can hence describe $\overline{\mathcal{O}_A}$ by the following formula:

$$\begin{aligned}
&\exists a_{1,1},\dots,a_{1,n},\dots,a_{n,1},\dots,a_{n,n}.\\
&\Bigg((x_1,\dots,x_n)\in K\wedge\Bigg(\bigwedge_{i=m_1+m_2+m_3+1}^{n}(x_i=0)\Bigg)\\
&\wedge(a_{i,j})_{i,j=1,\dots,n}\in\mathcal{M}\wedge\Bigg(\bigwedge_{i=1}^{n}\Bigg(y_i=\sum_{j=1}^{n}a_{i,j}x_j\Bigg)\Bigg)\Bigg).
\end{aligned}$$

This formula involves polynomials in n^2+2n variables. Their degrees are bounded by an expression in $d\tau+(n\tau)^{n^{O(1)}}$ and their coefficients are bounded in bitsize by an expression in $(n\tau)^{O(1)}$. By applying singly-exponential quantifier elimination (Theorem 2.12) we obtain that $\overline{\mathcal{O}_A}$ can be defined by a quantifier free formula involving polynomials of degree at most $(d\tau)^{n^{O(1)}}$ whose coefficients have bitsize at most $(d\tau)^{n^{O(1)}}$. The same bounds hold true for the description size of the set $S=\overline{\mathcal{O}_A}\times K$.

6.3.1.2 Description size of S'

Let us now bound the description size of the set S'. Recall that

$$\begin{aligned}
S'=\{(x,y,z)\in\mathbb{R}^{3n}\mid x\in K_{\mathrm{rec}}, y\in\overline{\mathcal{O}_A(x)}, z\in K,\\
\forall w\in\overline{\mathcal{O}_A(x)}.\,\mathrm{dist}_{\ell^2}(y,K)\geq\mathrm{dist}_{\ell^2}(w,K)\}.
\end{aligned}$$

For $w,y\in\overline{\mathcal{O}_A(x)}$ the predicate $\mathrm{dist}_{\ell^2}(y,K)\geq\mathrm{dist}_{\ell^2}(w,K)$ is expressed by the following first-order formula:

$$\exists m,n\forall v.\,(m,n,v\in K\wedge|y-v|\geq|y-m|\wedge|w-v|\geq|w-n|\wedge|y-m|\geq|w-n|)\,.$$

Let us write this as $\exists m,n\forall v.\Phi(y,w,m,n,v)$.

Hence, the predicate $\forall w\in\overline{\mathcal{O}_A(x)}.\,\mathrm{dist}_{\ell^2}(y,K)\geq\mathrm{dist}_{\ell^2}(w,K)$ can be written as

$$\Psi(x,y)=\forall w.\exists m,n.\forall v.\,\Big(w\notin\overline{\mathcal{O}_A(x)}\vee\Phi(y,w,m,n,v)\Big)\,.$$

Thus,

$$S'=\big\{(x,y,z)\in\mathbb{R}^{3n}\;\mid\;(x,y,z)\in S\wedge\Psi(x,y)\big\}\,.$$

The formula $\Psi(x, y)$ involves polynomials of degree at most $(d\tau)^{n^{O(1)}}$ and coefficients bounded in bitsize by $(d\tau)^{n^{O(1)}}$. Now, Theorem 2.12 yields a description size bound for S' of at most $(3n, (d\tau)^{n^{O(1)}}, (d\tau)^{n^{O(1)}})$.

6.3.1.3 Proving the recurrent bound for K_{rec}

By Theorem 2.14 we may obtain a lower-bounding function

$$\mathrm{LB}(n, d, \tau) \in 2^{(d\tau)^{n^{O(1)}}}$$

such that the minimum of the polynomial (6.2) over the set S' is bounded from below by $\frac{2\sqrt{n}}{\mathrm{LB}(n,d,\tau)}$. This means that for all $x \in K_{\text{rec}}$ there exists $y \in \overline{\mathcal{O}_A(x)}$ such that

$$\operatorname{dist}_{\ell^2}(y, K) > \frac{2\sqrt{n}}{\mathrm{LB}(n, d, \tau)}.$$

Now, consider the function

$$\mathrm{Rec}_0(n, d, \tau) = \max\{\mathrm{LB}(n, d, \tau), \mathrm{D}(n, d, \tau, \lceil\log(\mathrm{LB}(n, d, \tau))\rceil)\} \in 2^{(d\tau)^{n^{O(1)}}}.$$

We claim that the function Rec_0 has the property stated in Lemma 6.3. Let $x \in K_{\text{rec}}$. Let $y \in \overline{\mathcal{O}_A}(x)$ with $\operatorname{dist}_{\ell^2}(y, K) > \frac{2\sqrt{n}}{\mathrm{LB}(n,d,\tau)}$. Then by Lemma 6.4 there exists $t \leq \mathrm{Rec}_0(n, d, \tau)$ such that

$$\left\|A^t x - y\right\|_2 < \frac{1}{\mathrm{LB}(n, d, \tau)} \leq \frac{\sqrt{n}}{\mathrm{LB}(n, d, \tau)}.$$

For all $z \in K$ we then have

$$\left\|A^t x - z\right\|_2 > \|y - z\|_2 - \left\|A^t x - y\right\|_2 \geq \frac{\sqrt{n}}{\mathrm{LB}(n, d, \tau)} > \frac{\sqrt{n}}{\mathrm{Rec}_0(n, d, \tau)}.$$

This concludes the proof of Lemma 6.3.

6.3.2 Initial points in V_{rec}

It remains to extend the result to all initial points $x \in V_{\text{rec}}$. We first treat the special case where K_{rec} is empty.

Lemma 6.6. *There exists a function* $\mathrm{EmptyBound}(n, d, \tau) \in 2^{(d\tau)^{n^{O(1)}}}$ *with the following property:*

If K_{rec} is empty and $x \in V_{\text{rec}}$,

Then $\operatorname{dist}_{\ell^2}(x, K) > \sqrt{n}/\mathrm{EmptyBound}(n, d, \tau)$.

Proof. The function $\operatorname{dist}_{\ell^2}(\cdot, V_{\text{rec}})$ is the square root of a polynomial applied to the projection onto the vectorspace orthogonal to V_{rec}. By assumption,

$$\inf_{x \in K} \operatorname{dist}_{\ell^2}(x, V_{\text{rec}}) > 0.$$

Now, applying Theorem 2.14 to the square yields a lower bound of the desired shape. $\square$

We also need a technical lemma, which combines quantifier elimination with Lemma 2.15.

Lemma 6.7. *Assume that K_{rec} is non-empty. Let C be a positive integer. Further assume that K is contained in a ball of radius $\frac{1}{2} \cdot 2^{(d\tau)^{n^C}}$. Let $x \in V_{\text{rec}}$. If $\operatorname{dist}_{\ell^2}(x, K_{\text{rec}}) > 2^{-(d\tau)^{n^C}}$ then $\operatorname{dist}_{\ell^2}(x, K) > 2^{-(d\tau)^{n^{C+O(1)}}}$.*

Proof. Consider the function $\operatorname{dist}_{\ell^2}(x, K)$ on the set

$$L = \left\{ x \in V_{\text{rec}} \mid \|x\|_2 \leq 2^{(d\tau)^{n^C}} \wedge \operatorname{dist}_{\ell^2}(x, K_{\text{rec}}) \geq 2^{-(d\tau)^{n^C}} \right\}.$$

This set can be defined by the following first-order formula:

$$\begin{aligned}
&\forall (y_1, \ldots, y_n) \in \mathbb{R}^n . \forall (b_0, \ldots, b_m) \in \mathbb{R}^{n^C}. \\
&\left(\left((y_1, \ldots, y_n) \in K_{\text{rec}} \wedge b_0 = 2 \wedge b_{i+1} = b_i^{(d\tau)} \right) \rightarrow \|x\|_2^2 \leq b_m^2 \wedge \|x - y\|_2^2 \geq (1/b_m)^2 \right)
\end{aligned}$$

Applying quantifier elimination (Theorem 2.12), we find that the set L can be be defined by a quantifier-free formula involving polynomials of degree $(d\tau)^{O(n^C)}$ whose coefficients are bounded in bitsize by $(d\tau)^{O(n^C)}$.

It follows from Lemma 2.15 that every point in L has distance from K at least

$$1/\operatorname{Sep}(n, (d\tau)^{O(n^C)}, (d\tau)^{O(n^C)}) = 2^{\left(-(d\tau)^{O(n^C)}\right)^{n^{O(1)}}} = 2^{-(d\tau)^{n^{C+O(1)}}}.$$

Points in V_{rec} outside L are clearly at distance more than $2^{-(d\tau)^{n^C}}$ from K. $\square$

Let us now complete the proof of Proposition 6.2.

Recall that we wish to show existence of a function $\operatorname{Rec}(n, d, \tau) \in 2^{(\tau d)^{n^{O(1)}}}$ with the following property:

If every point $x \in K_{\text{rec}}$ escapes K under iterations of A then for all $x \in V_{\text{rec}}$ there exists $t \leq \operatorname{Rec}(n, d, \tau)$ such that

$$\operatorname{dist}_{\ell^2}(A^t x, K) > \frac{\sqrt{n}}{\operatorname{Rec}(n, d, \tau)}.$$

The function $\mathrm{Rec}_0(n,d,\tau)$ from the previous subsection is majorised by $2^{(d\tau)^{n^R}}$ for some constant R. Let $C \geq R+1$ be such that K is contained in a ball of radius $2^{(d\tau)^{n^C}}$. Such a constant exists by Theorem 2.13. Let D be a constant such that all $x \in V_{\mathrm{rec}}$ with $\mathrm{dist}_{\ell^2}(x, K_{\mathrm{rec}}) > 2^{-(d\tau)^{n^C}}$ satisfy $\mathrm{dist}_{\ell^2}(x,K) > 2^{-(d\tau)^{n^{C+D}}}$. The constant D exists thanks to Lemma 6.7.

Let

$$\mathrm{Rec}(n,d,\tau) := \max \left\{ \begin{array}{c} \mathrm{EmptyBound}(n,d,\tau), \\ \left\lceil \sqrt{n} 2^{(d\tau)^{n^{C+D}}} \right\rceil, \\ \left\lceil \frac{\sqrt{n}\,\mathrm{Rec}_0(n,d,\tau)}{\sqrt{n} - 2^{-(d\tau)^{n^C}}\,\mathrm{Rec}_0(n,d,\tau)} \right\rceil \end{array} \right\}.$$

Then, since we have chosen $C \geq R+1$, we have $\mathrm{Rec}(n,d,\tau) \in 2^{(d\tau)^{n^{O(1)}}}$.

Let us now show that Rec has the desired property. If K_{rec} is empty, then Rec has the desired property by construction. Now assume K_{rec} is non-empty.

Let $x \in V_{\mathrm{rec}}$. If $\mathrm{dist}_{\ell^2}(x,K) > 2^{-(d\tau)^{n^{C+D}}}$ then $\mathrm{dist}_{\ell^2}(x,K) > \sqrt{n}/\mathrm{Rec}(n,d,\tau)$. Assume that $\mathrm{dist}_{\ell^2}(x,K) \leq 2^{-(d\tau)^{n^{C+D}}}$. Then by Lemma 6.7 we have $\mathrm{dist}_{\ell^2}(x,K_{\mathrm{rec}}) \leq 2^{-(d\tau)^{n^C}}$. Choose $y \in K_{\mathrm{rec}}$ such that $\|x-y\|_2 < 2^{-(d\tau)^{n^C}}$. Then, by Lemma 6.3 there exists $t \leq \mathrm{Rec}_0(n,d,\tau)$ such that

$$\mathrm{dist}_{\ell^2}(A^t y, K) > \sqrt{n}/\mathrm{Rec}_0(n,d,\tau).$$

Since A is an isometry on V_{rec}, we have

$$\left\|A^t x - A^t y\right\|_2 = \|x-y\|_2 \leq 2^{-(d\tau)^{n^C}}.$$

It follows that

$$\mathrm{dist}_{\ell^2}(A^t x, K) > \sqrt{n}/\mathrm{Rec}_0(n,d,\tau) - 2^{-(d\tau)^{n^C}} > \sqrt{n}/\mathrm{Rec}(n,d,\tau).$$

6.4 The non-recurrent eigenspace bound

The next proposition constructs a bound for the subset $K_{\geq\varepsilon}$ of K containing points in K that are bounded away from V_{rec} by some $\varepsilon > 0$.

Proposition 6.8. *There exists an integer function* $\mathrm{NonRec}(n,d,\tau,P) \in 2^{(d\tau P)^{n^{O(1)}}}$ *with the following property:*

Let K be a compact semialgebraic set of description size at most (n,d,τ). Let $A \in \mathbb{A}^{n\times n}$ be a matrix in real Jordan normal form. Assume that the characteristic polynomial of A has rational coefficients whose bitsize is bounded by τ. Let P be a positive integer.

Then for all $x \in K_{\geq 2^{-P}}$ there exists $t \leq \mathrm{NonRec}(n,d,\tau,P)$ such that $A^t x \notin K_{\geq 2^{-P}}$.

Proof. We sketch an outline here. The rest of this section expands on this outline. For any point in $K_{\geq\varepsilon}$, by definition there exist coordinates (or pairs of coordinates if the corresponding eigenvalues are not real) whose contribution to the Jordan norm is greater than ε. Moreover, the Jordan norm of these coordinates does not stay constant under applications of A. If the norm of at least one such coordinate is increasing under applications of A, the orbit will eventually leave K, since K is compact. Theorem 2.13 yields an upper bound on the escape time.

Coordinates whose contribution to the norm is decreasing under applications of A will, after sufficiently many iterations, contribute less than ε. We establish a uniform upper bound on the number of iterations required to ensure this for all such coordinates. Combining this with the previous bound, we obtain a number N such that after at most N applications of A, every $x \in K_{\geq\varepsilon}$ has either escaped K, entered $K_{<\varepsilon} \cup K_{\text{rec}}$, or it remains in $K_{\geq\varepsilon}$ because it has a component whose contribution to the norm was initially smaller than ε, but grew beyond ε under iteration of A. In the last case, the point will grow in norm beyond the bound established in Theorem 2.13 and thus escape K after a further N applications of A. This yields a uniform bound on the number of iterations that are required for any point $x \in K_{\geq\varepsilon}$ to either leave K entirely or move into $K_{<\varepsilon} \cup K_{\text{rec}}$. □

The overall structure of this proof closely follows the one given in [33], which considers escape times for compact polytopes, and where the assumptions allow the authors to restrict the discussion to non-recurrent eigenvalues.

6.4.1 Jordan block bounds

Let J_k be a real Jordan block of multiplicity k corresponding to either a real eigenvalue Λ or a complex pair $\Lambda = a \pm ib$. We use t to denote positive integer time-steps.

$$J_k^t = \begin{bmatrix} \Lambda^t & t\Lambda^{t-1} & \binom{t}{2}\Lambda^{t-1} & \cdots & \binom{t}{k-1}\Lambda^{t-k+1} \\ 0 & \Lambda^t & t\Lambda^{t-1} & \cdots & \binom{t}{k-2}\Lambda^{t-k+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & t\Lambda^{t-1} \\ 0 & 0 & 0 & \cdots & \Lambda^t \end{bmatrix}.$$

where Λ can be considered $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ or a scalar quantity depending on the type of eigenvalue.

Thus we see that $x(t) := J_k^t x$ can be written component-wise as

$$x_j(t) = \sum_{i=j}^{k} \binom{t}{i-j} \Lambda^{t-(i-j)} x_i,$$

where the 'components' $x_1 \dots x_k$ refer to scalars (if the eigenvalue is real) or vectors $(x_j^{(r)}, x_j^{(i)})$ if the eigenvalue is complex.

We will use absolute value signs to denote the Jordan norm $\|\cdot\|_J$ of a component x_j. Note this is the same as the ℓ_2 norm since we consider only one component.

Then we have the following three lemmas, where we case split on the modulus of the eigenvalue Λ, which we notate as γ, to obtain block-wise bounds on escape times for non-recurrent eigenspaces:

These lemmas have the same structure as those in [33], however for convenience we present the discrete case in full here, as [33] formulates them in the continuous case.

Lemma 6.9 (Polynomially expanding Jordan block - $\gamma = 1$)**.** *Let C be a positive number such that K is contained in the ℓ^2-ball centred at 0 with radius C. Let $x(t) := J_k^t x$ where x is an initial point in K. Assume there exists $j \geq 2$ with $|x_j| > \varepsilon$ in the Jordan block. Let $N = \frac{1}{k}\left(\frac{k^2 C}{\varepsilon}\right)^{2^{k-1}}$. Then there exists j and $t \leq N$ with $|x_j(t)| > C$.*

Proof. Given the set of equations

$$x_j(t) = \sum_{i=j}^{k} \binom{t}{i-j} \Lambda^{t-(i-j)} x_i,$$

we want an N such that there exists j such that $|x_j(N)| > C$.

Let $j_1 \geq 2$ be the smallest j such that $|x_{j_1}| > \varepsilon$ (we are given that such a component exists). Note that we require $j_1 \geq 2$ because x_1 is actually in V_{rec} since $\gamma = 1$.

Consider the component

$$x_{j_1-1}(t) = \sum_{i=j_1-1}^{k} \binom{t}{i-j_1+1} \Lambda^{t-(i-j_1+1)} x_i.$$

We set $N_{j_1} = kC/\varepsilon$. Observe that (note $\gamma = 1$)

$$|x_{j_1-1}(N_{j_1})| \geq |(kC/\varepsilon)x_{j_1}| - |x_{j_1-1}| - \sum_{i=j_1+1}^{k} \left| \binom{N_{j_1}}{(i-j_1+1)} \Lambda^{t-(i-j_1+1)} x_i \right|$$

Since the first term is larger than kC and the second term is smaller than C, the only way $|x_{j_1-1}(N_{j_1})|$ could be less than C (and thus not escape K) is if one of the later terms is larger than C. Let j_2 be the highest index such that $\left| x_{j_2} \binom{N_{j_1}}{j_2-j_1+1} \right| \geq C$.

Note that $j_2 > j_1$. We now have a lower bound on a higher index coefficient, namely $|x_{j_2}| \binom{N_{j_1}}{j_2 - j_1 + 1} \geq C$. We now repeat the process with the component

$$x_{j_2-1}(t) = \Lambda^t x_{j_2-1} + \binom{t}{1} \Lambda^{t-1} x_{j_2} + \sum_{i=j_2+1}^{k} \binom{t}{i - j_2 + 1} \Lambda^{t-(i-j_2+1)} x_i$$

We have $|x_{j_2}| \binom{N_{j_1}}{j_2 - j_1 + 1} \geq C$ thus setting $N_{j_2} > k \binom{N_{j_1}}{j_2 - j_1 + 1}$ ensures that $|N_{j_2} x_{j_2}| > kC$.

Continuing this process, we will either find a component that escapes the set or move on to a component with higher index, which can happen at most $k - 1$ times, because we have the constraints $j_1 \geq 2, \forall m, j_m \leq k$, and $j_m > j_{m-1}$. This gives us a recursive definition for the bound, which is

$$N_{j_m} > k \binom{N_{j_{m-1}}}{j_m - j_{m-1} + 1}$$

We wish to find an upper bound on $N = N_M$, the time by which we are guaranteed that at least one component escapes, subject to the constraints $j_1 \geq 2, j_M \leq k$, and $j_m > j_{m-1}$. We can solve the recursive inequality by weakening it (since we only need an upper bound on N) to

$$N_{j_m} > (k N_{j_{m-1}})^{j_m - j_{m-1} + 1}.$$

Note that pulling the constant k into the exponentiated part is valid because $j_m - j_{m-1} + 1 > 2$ always. Setting $S_{j_m} = kN_{j_m}$, we get $S_{j_m} > S_{j_{m-1}}^{j_m - j_{m-1}+1}, S_{j_1} = k^2 C/\varepsilon$, which reduces to

$$S_{j_M} > \left(\frac{k^2 C}{\varepsilon}\right)^{\prod_{m=2}^{M} (j_m - j_{m-1} + 1)}$$

The term $\prod_{m=2}^{M} (j_m - j_{m-1} + 1)$ is maximised when for all m, $j_m = j_{m-1} + 1$, thus in the worst case we have

$$S_{j_M} > \left(\frac{k^2 C}{\varepsilon}\right)^{2^{k-1}}$$

Thus we have a bound for a modulus-1-eigenvalue component to escape, which is

$$\boxed{N = \frac{1}{k} \left(\frac{k^2 C}{\varepsilon}\right)^{2^{k-1}}.}$$

□

Lemma 6.10 (Exponentially shrinking Jordan block: $\gamma < 1$)**.** *Let $C \geq n$ be a positive number such that K is contained in the ℓ^2-ball centred at 0 with radius C. Let $x \in K$. Let ε be a positive real number. Let $N = \frac{4k}{\log(1/\gamma)} \log\left(\frac{2k}{\log(1/\gamma)}\right) + \frac{2\log(kC/\varepsilon)}{\log(1/\gamma)}$. Then there exists $t \leq N$ with $|x_j(t)| < \varepsilon$ for all j in the block.*

Proof. If $\gamma = 0$ then we have $x_j(t) = 0$ for all j in the block, so we may assume $\gamma > 0$ for the rest of the proof.

We have the equations

$$x_j(t) = \sum_{i=j}^{k} \binom{t}{i-j} \Lambda^{t-(i-j)} x_i,$$

For any j, we have the result

$$|x_j(t)| \leq \sum_{i=j}^{k} \binom{t}{i-j} \cdot |\Lambda^{t-(i-j)} x_i| < kC(et/k)^k \gamma^{t-k},$$

where the second inequality is obtained via standard bounds on binomial coefficients

In order to have $|x_j(t)| < \varepsilon$, it is enough to have $kC(et/k)^k\gamma^{t-k} < \varepsilon$, which is equivalent (after weakening slightly by dropping irrelevant terms)to $t > \frac{k}{\log(1/\gamma)} \log t + \frac{\log(kC/\varepsilon)}{\log(1/\gamma)}$.

Here we need a small technical lemma.

Lemma 6.11 ([88, Lemma A.1, Lemma A.2])**.** *Suppose $a \geq 1$ and $b > 0$, then $t \geq a \log t + b$ if $t \geq 4a \log(2a) + 2b$.*

Applying this lemma we get a bound on N such that for all $j \leq k$, $x_j(N) < \varepsilon$, namely

$$\boxed{N \leq \frac{4k}{\log(1/\gamma)} \log\left(\frac{2k}{\log(1/\gamma)}\right) + \frac{2\log(kC/\varepsilon)}{\log(1/\gamma)}.}$$

□

Lemma 6.12 (Exponentially expanding Jordan block - $\gamma > 1$)**.** *Let C be a positive number such that K is contained in the ℓ^2-ball centred at 0 with radius C. Let $x(t) := J_k^t x$ where x is an initial point in K. Assume there exists j with $|x_j| > \varepsilon$ in the Jordan block. Let $N = 2^{k-1}\left(\frac{\log(kC/\varepsilon)}{\log\gamma} + \frac{4k}{\log\gamma}\log\left(\frac{2k}{\log\gamma}\right) + \frac{k}{\log\gamma}\right)$. Then there exists j and $t \leq N$ with $|x_j(t)| > C$.*

Proof. This proof combines ideas from the previous two cases.

By assumption, there exists $j_1 \geq 1$ such that $|x_{j_1}| > \varepsilon$. Consider the component

$$x_{j_1}(t) = \sum_{i=j_1}^{k} \binom{t}{i-j_1} \Lambda^{t-(i-j_1)} x_i.$$

Set $N_{j_1} = \frac{\log(kC/\varepsilon)}{\log\gamma}$ and observe that

$$|x_{j_1}(N_{j_1})| \geq |\Lambda^{\frac{\log(kC/\varepsilon)}{\log\gamma}} x_{j_1}| - \sum_{i=j_1+1}^{k} \left| \binom{N_{j_1}}{i-j_1} \Lambda^{N_{j_1}-(i-j_1)} x_i \right|.$$

Since the first term is larger than kC, the only way $|x_{j_1}(N_{j_1})|$ can be less than C (and thus not escape the set) is if one of the later terms is larger than C. Let j_2 be the highest index such that $\left|\binom{N_{j_1}}{j_2-j_1}\Lambda^{N_{j_1}-(j_2-j_1)}x_{j_2}\right| \geq C$. Note that $j_2 > j_1$. We now have a lower bound on a higher index coefficient, namely $\binom{N_{j_1}}{j_2-j_1}\gamma^{N_{j_1}-(j_2-j_1)}|x_{j_2}| \geq C$. Now we repeat the process with the component

$$x_{j_2}(t) = \Lambda^t x_{j_2} + \sum_{i=j_2+1}^{k} \binom{t}{i-j_2}\Lambda^{t-(i-j_2)}x_i$$

We want $|\Lambda^t x_{j_2}| > kC$, so it is enough to set

$$\gamma^{N_{j_2}} > k\binom{N_{j_1}}{j_2-j_1}\gamma^{N_{j_1}-(j_2-j_1)}.$$

Similarly to the previous lemma, this gives us a recursive definition for the bound, which is

$$N_{j_m} - N_{j_{m-1}} > \frac{\log\binom{N_{j_{m-1}}}{j_m-j_{m-1}}}{\log\gamma} + \frac{\log k}{\log\gamma} - (j_2 - j_1).$$

We may weaken this (using standard binomial bounds) to

$$N_{j_m} - N_{j_{m-1}} > \frac{k\log N_{j_{m-1}}}{\log\gamma} + \frac{k}{\log\gamma}.$$

Let us assume $N_{j_{m-1}} > \frac{k\log N_{j_{m-1}}}{\log\gamma} + \frac{k}{\log\gamma}$. By the same technical lemma (Lemma 6.11) this will hold if $N_{j_{m-1}} > \frac{4k}{\log(\gamma)}\log\left(\frac{2k}{\log(\gamma)}\right) + \frac{k}{\log(\gamma)}$. Then we may simplify the inequality to

$$N_{j_m} > 2N_{j_{m-1}},$$

which is easily solved to get

$$N_M > 2^{M-1}N_{j_1}.$$

As $M \leq k$, setting $N_{j_1} = \frac{\log(kC/\varepsilon)}{\log\gamma} + \frac{4k}{\log(\gamma)}\log\left(\frac{2k}{\log(\gamma)}\right) + \frac{k}{\log(\gamma)}$ gives us a bound for a component to escape, which is

$$\boxed{N \leq 2^{k-1}\left(\frac{\log(kC/\varepsilon)}{\log\gamma} + \frac{4k}{\log\gamma}\log\left(\frac{2k}{\log\gamma}\right) + \frac{k}{\log\gamma}\right).}$$

$\square$

6.4.2 Combining the block-wise bounds

We now complete the proof of Proposition 6.8.

Recall that we wish to show there exists an integer function $\text{NonRec}(n, d, \tau, P) \in 2^{(d\tau P)^{n^{O(1)}}}$ with the property that for all $x \in K_{\geq 2^{-P}}$ there exists $t \leq \text{NonRec}(n, d, \tau, P)$ such that $A^t x \notin K_{\geq 2^{-P}}$

We will show there exists a positive integer L such that

$$\text{NonRec}(n, d, \tau, P) := 2^{P \cdot 2^n} \cdot 2^{(\tau \cdot d)^{Ln}} \in 2^{(d\tau P)^{n^{O(1)}}}$$

satisfies the conditions of Proposition 6.8.

Set $\varepsilon = 2^{-P}$. Define a partition of K into $K_{\text{rec}}, K_{<\varepsilon}$ and $K_{\geq\varepsilon}$ as described in section 6.2. We need to construct $N_{\geq\varepsilon}$ such that for all $x \in K_{\geq\varepsilon}$ there exists $t \leq N_{\geq\varepsilon}$ such that $A^t x \notin K_{\geq\varepsilon}$, which is to say, we have either escaped the set K completely or moved into $K_{<\varepsilon} \cup K_{\text{rec}}$.

Since $x \in K_{\geq\varepsilon}$, we start with at least one component greater than ε in Jordan norm. This allows us to leverage the block-wise bounds.

Let $N_{\geq\varepsilon} = 2\max_\gamma\{N_\gamma\}$, where N_γ ranges over the possible bounds depending on the size of the eigenvalues of A. Within a time $N_{\geq\varepsilon}/2 = \max_\gamma\{N_\gamma\}$, thanks to the analysis of the block-wise bounds, there are three possibilities:

- the orbit has increased in size beyond C and has thus left K. This occurs if there was a component associated to an expanding eigenvalue that was larger than ε in Jordan norm;
- all components are now smaller than ε, thus leaving $K_{\geq\varepsilon}$ (by entering $K_{<\varepsilon}$);
- some component corresponding to an expanding eigenvalue which was originally less than ε has become greater than ε. In this case, waiting another $N_{\geq\varepsilon}/2$ amount of time puts the trajectory in the first case, ensuring it escapes.

Thus in all cases the trajectory has escaped by time $N_{\geq\varepsilon}$.

We now explicitly compute $N_{\geq\varepsilon}$ by using the description size of K and A to bound the three possibilities, namely

$$\boxed{N_\gamma = \frac{4k}{\log(1/\gamma)} \log\left(\frac{2k}{\log(1/\gamma)}\right) + \frac{2\log(kC/\varepsilon)}{\log(1/\gamma)}.}$$

(shrinking eigenvalue γ),

$$\boxed{N_\gamma = 2^{k-1}\left(\frac{\log(kC/\varepsilon)}{\log\gamma} + \frac{4k}{\log\gamma}\log\left(\frac{2k}{\log\gamma}\right) + \frac{k}{\log\gamma}\right).}$$

(exponentially expanding eigenvalue γ) and

$$\boxed{N_\gamma = \frac{1}{k}\left(\frac{k^2C}{\varepsilon}\right)^{2^{k-1}}.}$$

(modulus 1 eigenvalue), where k is the multiplicity of the Jordan block and thus $k \leq n$, the dimension.

We now compute bounds on $\log\gamma$ and C in terms of K and A.

Bounding $\log\gamma$**:** Let τ be a bound on the bitsize of the coefficients of the minimal polynomial of the eigenvalues of A as well as a bound on the total bitsize of K.

Using the fact that $x/2 < \log(1+x) < x$ for $|x| << 1$, Lemma 2.3 yields a constant R satisfying $\frac{1}{\log\gamma} < 2^{(\tau n)^R}$.

Bounding C**:** Let d bound the degree of the polynomials defining K. Then from Theorem 2.13 we have the existence of a constant S satisfying $C < 2^{(\tau\cdot d)^{S(n+1)}}$.

Plugging these bounds into the iteration bounds from the previous lemmas, and overapproximating for simplicity (we may trivially assume $\varepsilon < 1$), we finally get the following bound: With a new constant Q based on R and S, we have

$$N_{\geq\varepsilon} \leq \left(\frac{1}{\varepsilon}\right)^{2^n} \cdot 2^{(2^n\cdot(\tau\cdot d)^{S(n+1)})} + (2\cdot\tau\cdot d)^{(\tau n)^Q} + \log(1/\varepsilon)\cdot 2^{(\tau n)^Q}.$$

We can simplify this still further by amalgamating terms. Letting L be a new constant appropriately larger than S and Q, we set

$$\boxed{N_{\geq\varepsilon} = \left(\frac{1}{\varepsilon}\right)^{2^n} \cdot 2^{(\tau\cdot d)^{Ln}}}.$$

Observe that the time to leave $K_{\geq\varepsilon}$ is doubly exponential in the dimension, singly exponential in the rest of the input data, and inverse polynomial in ε.

Now $\mathrm{NonRec}(n, d, \tau, P) := N_{\geq 2^{-P}} = 2^{P\cdot 2^n} \cdot 2^{(\tau\cdot d)^{Ln}}$ satisfies the requirements of Proposition 6.8.

6.5 Putting it all together

In the previous two sections, we successively showed how to establish a bound on the escape time for an instance (A, K) when the orbit remains in the recurrent eigenspace and how the orbit behaves when it starts away from the recurrent eigenspace. In this section, we show how to combine both results in order to establish an escape bound for any starting point in K. This will thus prove Theorem 6.1.

Let (A_0, K_0) be an instance of the compact escape problem, where $K_0 \subseteq \mathbb{R}^n$ is a compact semialgebraic set of description size at most (n_0, d_0, τ_0) and $A_0 \in \mathbb{Q}^{n\times n}$ is a square matrix with rational entries whose bitsize is bounded by τ_0. Assume that every point $x \in K_0$ escapes K_0 under iterations of A_0.

Apply Lemma 2.19 to convert the instance (A_0, K_0) into an equivalent instance (A, K) such that $A \in (\overline{\mathbb{Q}} \cap \mathbb{R})^{n\times n}$ is in real Jordan normal form. Then the set K has description size at most (n, d, τ), where $n = 2n_0$, $d = n_0 d_0$, and $\tau = (n_0\tau_0 d_0)^{C_\tau}$ for some absolute constant C_τ. By construction, the characteristic polynomial of A has rational coefficients of bitsize at most τ.

Let Rec be the function from Proposition 6.2. Let $\varepsilon = \frac{1}{\mathrm{Rec}(n,d,\tau)}$ and $N_{\mathrm{rec}} = \mathrm{Rec}(n, d, \tau)$. Let $x \in K$. If $x \in K_{\mathrm{rec}}$ then x escapes within N_{rec} steps. Suppose that $x \in K_{<\varepsilon}$.

Then there are two possibilities:

1. We have $A^t x \notin K_{\geq\varepsilon}$ for all $t \leq N_{\mathrm{rec}}$.

2. We have $A^t x \in K_{\geq\varepsilon}$ for at least one $t \leq N_{\mathrm{rec}}$.

In the first case, the orbit of x remains close to V_{rec} for long enough that we can rely on Proposition 6.2. Indeed, let x_0 denote the orthogonal projection of x onto V_{rec}. Let $t \leq N_{\mathrm{rec}}$ be such that $\mathrm{dist}_{\ell^2}(A^t x_0, K) > \sqrt{n}\varepsilon$. Since $A^t x \notin K_{\geq\varepsilon}$, we have $\|A^t x - A^t x_0\|_J < \varepsilon$, so that $\|A^t x - A^t x_0\|_2 < \sqrt{n}\varepsilon$. Let $y \in K$. Then

$$\left\|A^t x - y\right\|_2 \geq \left\|A^t x_0 - y\right\|_2 - \left\|A^t x - A^t x_0\right\|_2 > \sqrt{n}\varepsilon - \sqrt{n}\varepsilon = 0.$$

Thus, x escapes K under iterations of A.

In the second case, let t_1 be such that $A^{t_1} x \in K_{\geq\varepsilon}$. Let NonRec be the function from Proposition 6.8. Let $N_{\geq\varepsilon} = \mathrm{NonRec}(n, d, \tau, \lceil\log(1/\varepsilon)\rceil)$. By Proposition 6.8 there exists $t_2 \leq N_{\geq\varepsilon}$ such that $A^{t_2} A^{t_1} x$ is contained either in $K_{<\varepsilon} \cup K_{\mathrm{rec}}$ or in the complement of K. In the latter case we are done. In the former case we apply the initial case distinction: either for all $t \leq N_{\mathrm{rec}}$ we have $A^t A^{t_2} A^{t_1} x \notin K_{\geq\varepsilon}$ or we have $A^{t_3} A^{t_2} A^{t_1} x \in K_{\geq\varepsilon}$ for at least one $t_3 \leq N_{\mathrm{rec}}$. Once again, in the first case, the point has escaped. By repeating this reasoning, we construct a (finite or infinite) sequence $t_1, t_2, \ldots$ such that $t_i \leq N_{\mathrm{rec}}$ if i is odd and $t_i \leq N_{\geq\varepsilon}$ if i is even and

$$A^{t_s} \cdot \cdots \cdot A^{t_1} x \in \begin{cases} K_{<\varepsilon} \cup K_{\mathrm{rec}} & \text{if } s \text{ is even,} \\ K_{\geq\varepsilon} & \text{if } s \text{ is odd.} \end{cases}$$

6.5.1 Bounding the sequence of "flips"

We claim that the sequence $t_1, t_2, \dots$ is finite and contains at most n^3 elements. Intuitively, the exponential-polynomial defined by $A^t x$ can "flip" in size above and below the threshold only due to the oscillations of the polynomial part, and these are bounded in number by the polynomial degree, which is at most n.

Consider a real Jordan block of A of size $m \leq n$ associated to the eigenvalue Λ. Denote by x_J the orthogonal projection of x onto the dimensions associated with this block.

Assume first that Λ is a real eigenvalue (as opposed to a 2×2 block representing a complex eigenvalue). If $\Lambda = 0$, then clearly $\left\|J^k x_J\right\|_J$ is monotonically decreasing. Thus, assume in the sequel that $\Lambda \neq 0$.

Let $j \in \{1, \dots, m\}$. The $m - j + 1$'th component of the vector $J^k x_J$, viewed as a function of t, is an exponential polynomial $E_j(t) = \Lambda^t P(t)$, where $P \in \mathbb{R}[z]$ is a real polynomial of degree $j - 1$. Consider the real function

$$(E_j(\cdot))^2 \colon \mathbb{R} \to \mathbb{R},\ (E_j(t))^2 = |\Lambda|^{2t} |P(t)|^2.$$

This function is differentiable in t with derivative

$$\tfrac{d}{dt}(E_j(t))^2 = \Lambda^{2t} \left(\log(\Lambda^2)(P(t)^2) + 2P(t)P'(t)\right).$$

This derivative vanishes if and only if the factor $(\log(\Lambda^2)(P(t)^2) + 2P(t)P'(t))$ vanishes. This factor is a polynomial of degree $2j - 2$, so that it has at most $2j - 2$ real zeroes. It follows that there exist numbers $t_{j,1}, \dots, t_{j,m_j}$ with $m_j \leq 2j - 2$ such that the function $(E_j(t))^2 - \varepsilon^2$ does not change its sign in any of the open intervals

$$(0, t_{j,1}), (t_{j,1}, t_{j,2}), \dots, (t_{j,m_j-1}, t_{j,m_j}), (t_{j,m_j}, +\infty).$$

Thus, the norm $\|J^t x_J\|_J$ changes from smaller than ε to bigger than ε at most

$$\sum_{j=1}^{m}(2j-2) = 2\sum_{j=1}^{m} j - 2m = (m+1)m - 2m = m^2 - m$$

times.

The case where Λ represents a complex eigenvalue λ is similar. However, we now consider the evolution of the two coordinates corresponding to one Λ-block simultaneously.

For $j \in \{1, \dots, m\}$, write $E_j(t)$ for the $m - j + 1$'th component of the vector $J^t x_J$, viewed as a function of t. We have for all $j \in \{1, \dots, m/2\}$ that the function

$$F_j(t) = (E_{2j}(t))^2 + (E_{2j-1}(t))^2$$

is an exponential polynomial $F_j(t) = |\lambda|^t P_j(t)$, where $P_j \in \mathbb{R}[z]$ is a real polynomial of degree $j-1$. Therefore, exactly as in the case where Λ is a real eigenvalue, the derivative of F_j vanishes at most $2j-2$ times. From which we can deduce that the norm $\|J^t x_J\|_J$ crosses the ε-threshold at most $m^2 - m$ times.

Estimating generously, we have at most n Jordan blocks of size at most n, each of which crosses the ε-threshold at most $n^2 - n$ times. In total, we cross the threshold at most $n^3 - n^2$ times. The total escape bound is hence $n^3 \max\{N_{\mathrm{rec}}, N_{\geq\varepsilon}\}$. By the same argument, the same escape bound holds true when the initial point x lies in $K_{\geq\varepsilon}$.

6.5.2 The final bound

Substituting the constants N_{rec}, $N_{\geq\varepsilon}$, n, d, and τ with their definitions, we obtain the upper bound

$$\begin{aligned}&\mathrm{CompactEscape}(n_0, d_0, \tau_0) = \\ &\quad (2n_0)^3 \max\Big\{ \mathrm{Rec}\left(2n_0, n_0 d_0, (n_0 d_0 \tau_0)^{C_\tau}\right), \\ &\qquad\qquad \mathrm{NonRec}\left(2n_0, n_0 d_0, (n_0 d_0 \tau_0)^{C_\tau}, \log\left\lceil \mathrm{Rec}\left(2n_0, n_0 d_0, (n_0 d_0 \tau_0)^{C_\tau}\right)\right\rceil\right) \Big\}.\end{aligned}$$

One easily verifies that $\mathrm{CompactEscape}(n, d, \tau) \in 2^{(d\tau)^{n^{O(1)}}}$ as claimed.

6.6 A matching lower bound on escape time

In Theorem 6.1 we established a uniform upper bound on the escape time for all positive instances of the Compact Escape Problem. Our bound is doubly exponential in the ambient dimension and singly exponential in the rest of the data. We will now show that this bound cannot be significantly improved by showing that a doubly exponential bound cannot be avoided even for purely rotational systems.

6.6.1 Rotational example

The idea of this example is to use the dimensions iteratively to create a circle with a double-exponentially tiny hole which is the only escape.

Example. For $(n, d, \tau) \in \mathbb{N}^3$, let $K_{(n,d,\tau)} \subseteq \mathbb{R}^{n+2}$ be the set of all points $(x, y, u_1, \ldots, u_n)$ satisfying the (in)equalities: $x^2 + y^2 = 1, u_1 = 2^{-\tau}$, $(x-1)^2 + y^2 \geq u_n$ and for $1 \leq i \leq n-1$, $u_{i+1} = (u_i)^d$.

Hence, $K_{(n,d,\tau)} = \left(S^1 \setminus B\left((1,0), 2^{-\tau d^{n-1}}\right)\right) \times \left\{\left(2^{-\tau}, 2^{-\tau d}, \ldots, 2^{-\tau d^{n-1}}\right)\right\}$, where $S^1 \subseteq \mathbb{R}^2$ is the unit circle. Let $a = \frac{3}{5}$, $b = \frac{4}{5}$. Let

$$A_{(n,d,\tau)} = \begin{pmatrix} a & -b & 0 \\ b & a & 0 \\ 0 & 0 & I_n \end{pmatrix}$$

where I_n is the $n \times n$ identity matrix. It is easy to see that the complex number $\frac{3}{5} + i\frac{4}{5}$ has modulus 1 and is not a root of unity. It follows from Dirichlet's theorem on simultaneous Diophantine approximation that the orbit of A is equal to $S^1 \times \left\{\left(2^{-\tau}, 2^{-\tau d}, \ldots, 2^{-\tau d^{n-1}}\right)\right\}$, so that every initial point escapes under A.

We claim that there exists a point $x \in K_{(n,d,\tau)}$ that requires $2^{\tau d^{n-1}}$ steps to escape. Indeed, let $x_0 \in K_{(n,d,\tau)}$ be an arbitrary initial point. Consider the orbit $x_t = A^t x_0$. Let $N < 2^{\tau d^{n-1}}$. By the pigeonhole principle, the finite set of points $x_0, \ldots, x_N$ contains at least one consecutive pair of points x_i, x_j on the circle such that the points x_i and x_j are joined by an arc of the circle of length strictly greater than $2/N$. It follows that we can ensure that none of the points $x_1, \ldots, x_N$ is outside of $K_{(n,d,\tau)}$ by applying a suitable planar rotation to all points. Since all planar rotations commute, there exists for each angle θ an initial point $x_\theta \in S^1 \times \left\{\left(2^{-\tau}, 2^{-\tau d}, \ldots, 2^{-\tau d^{n-1}}\right)\right\}$, such that the orbit of x_θ under A is equal to the orbit of x_0 under A rotated by θ. This proves the claim.

6.6.2 Expanding example

Now we present another example where the doubly exponential bound comes from the size of the set we define.

Example. The construction of our first family of instances $(K_{(n,d,\tau)}, A_{(n,d,\tau)})_{(n,d,\tau)\in\mathbb{N}^3}$ relies on the fact that one can define a compact semialgebraic set whose size is doubly-exponential in the ambient dimension.

For $(n,d,\tau) \in \mathbb{N}^3$, define $K_{(n,d,\tau)} \subseteq \mathbb{R}^{n+1}$ as the set of all points $(x_1, \ldots, x_n, x_u)$ satisfying the (in)equalities:

$$\begin{aligned} &x_u = 1, \\ &x_1 = 2^\tau, \\ &\text{For } 1 \le i \le n-2,\ x_{i+1} = x_i^d, \\ &0 \le x_n \le x_{n-1}^d. \end{aligned}$$

Thus, a point $x \in \mathbb{R}^{n+1}$ belongs to $K_{(n,d,\tau)}$ if and only if it is of the form $\left(2^\tau, 2^{\tau d}, \ldots, 2^{\tau d^{n-2}}, y, 1\right)$ where $y \in \left[0, 2^{\tau d^{n-1}}\right]$.

We now define $A_{(n,d,\tau)}$ to be the matrix which only adds 1 (through the coefficient x_u) to the penultimate coordinate:

$$A_{(n,d,\tau)} = \begin{pmatrix} 1 & 0 & \dots & & \\ 0 & 1 & \dots & & \\ \vdots & \vdots & \ddots & & \\ & & & 1 & 1 \\ & & & 0 & 1 \end{pmatrix}$$

Therefore, given an initial point $x_0 \in K_j$, we have that $x_t = A^t_{(n,d,\tau)} x_0 = x_0 + (0, \dots, 0, t, 0)$. This sequence obviously escapes. The point $(2^\tau, 2^{\tau d}, \dots, 2^{\tau d^{n-2}}, 0, 1) \in K_{(n,d,\tau)}$ requires $2^{\tau d^{n-1}+1}$ iterations to escape. This is doubly exponential in the ambient dimension and singly exponential in the rest of the data.

6.7 Conclusion

In this chapter we have shown upper and lower bounds for the escape time of a linear dynamical system over a compact semialgebraic set that are doubly exponential in the ambient dimension. Both our lower bound examples rely on using a repeated squaring of polynomials in order to make the set itself double-exponentially large or small.

This raises the question of what the optimal bound is for a compact polytope, where this strategy would not be available. Previous work has shown a doubly exponential upper bound for the escape time of a continuous linear dynamical system over a compact polytope [33]. No lower bound was given.

The double exponential occurs in the analogue of Lemma 6.9, which bounds how slowly a single variable polynomial can grow in terms of a lower bound on *one* coefficient. The proof was not optimized, and it is very plausible that a more careful analysis could yield a singly exponential bound.

Chapter 7

Discussion and Open Problems

In this thesis we have studied reachability for matrices and pseudo-orbits, and escape for linear dynamical systems. Each chapter-specific conclusion discusses possible extensions to those results. We now discuss some common themes.

We showed that the Non-Negative Membership Problem is undecidable in general, but decidable for commutative matrix semigroups subject to Schanuel's conjecture. The use of Schanuel's conjecture arises because we reduce the commutative case of the Non-Negative Membership Problem to the decision problem for the first-order theory of real-closed fields with exponentiation. However in the diagonalisable case, we only need Schanuel's conjecture to decide whether a rational function in logarithms of algebraic numbers is zero.

We have shown decidability of the pseudo-Skolem problem in all dimensions, and decidability of pseudo-reachability for arbitrary semialgebraic sets when the matrix is diagonalisable.

A connection that emerges is the relative tractability of questions involving diagonalisable matrices. Indeed our third topic also follows this pattern: while the Escape Problem (deciding whether a set is an escape set for a given matrix) is open for general non-compact semialgebraic sets, unpublished work with Lefaucheux and Neumann indicates that it is decidable for diagonalisable systems (via a reduction to Ultimate Positivity for simple linear recurrences, which is decidable [75]).

It is well known [74] that decidability of the Positivity Problem or Ultimate Positivity Problem would lead to advances in determining the Lagrange constants of a large class of real numbers, an open problem in Diophantine approximation which can be seen as a number-theoretic barrier to further progress. However this reduction crucially relies on the nondiagonalisability of the matrix.

Thus it is interesting to consider barriers to decidability in the diagonalisable case and the implications of further progress.

Recent work [54] has shown that solving diagonalisable Positivity would have implications beyond reachability: it would allow deciding whether the trajectory of a diagonalisable LDS satisfies arbitrary omega-regular properties (defined on semialgebraic predicates). However, this implication does not hold for general Positivity and general LDS.

The proof of decidability of Ultimate Positivity for simple linear recurrences uses a nonconstructive lower bound on sums of S-units based on the Subspace Theorem [75]. An effective version of the Subspace Theorem would be a direct, if very difficult, route to solving diagonalisable Positivity.

Results on deciding diagonalisable Positivity for dimension up to 9 have used Baker's theorem to bound the orbit away from a finite set of points [76]. Further progress to dimension 10 and beyond is bottlenecked by the fact that one would need an analogue of Baker's theorem that enabled one to bound the orbit away from a *curve*. Note also that decidability in dimension 14 would solve the last remaining open case of Skolem at dimension 5.

Separately, we could drop our reliance on Schanuel's conjecture in the diagonalisable commuting case of the Non-Negative Membership Problem if it was possible to bound a nonzero rational function in logarithms of algebraic numbers away from zero. This would be a very significant extension of Baker's theorem, subsuming the four exponentials conjecture.

Thus, while some of the barriers to Positivity do not apply, solving the diagonalisable version still requires major advances in transcendental number theory.

7.1 Open Problems for Future Work

Returning to problems that (hopefully) do not require such advances, we list some open problems for future work that we think are interesting and relatively tractable.

1. Prove that Eventual Nonnegativity for a single matrix is in PTIME. Characterising the nonnegative powers of a matrix requires splitting into non-degenerate sequences, which might cause an exponential blowup. We speculate that for Eventual Nonnegativity one can avoid this by proving that degenerate matrices with negative entries cannot be eventually nonnegative. (See Section 3.2)

2. Prove that the Positive Membership Problem is undecidable. (See Section 3.5.1)

3. Characterise the complexity of the pseudo-Skolem problem. (See Theorem 4.23)

4. Prove decidability (or hardness) for pseudo-reachability for non-diagonalisable systems. As mentioned in Section 5.6, the simplest tricky case is to characterise the set-summation $\sum_{i=0}^{n-1} M^i B$ where $M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and B is a unit ball.

5. Prove that the escape time for a compact *polytope* is singly exponential in the dimension. This would be a significant improvement over the doubly exponential bound given in [33].

Bibliography

[1] S. Akshay, Hugo Bazille, Blaise Genest, and Mihir Vahanwala. On Robustness for the Skolem and Positivity Problems. In *Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 5:1–5:20, 2022.

[2] S. Akshay, S. Chakraborty, and D. Pal. On eventual non-negativity and positivity for the weighted sum of powers of matrices. In *IJCAR*, 2022.

[3] Bowen Alpern and Fred B. Schneider. Defining liveness. *Information Processing Letters*, 21(4):181–185, 1985.

[4] L. Babai, R. Beals, J.-Y. Cai, G. Ivanyos, and E. M. Luks. Multiplicative equations over commuting matrices. In *SODA*, pages 498–507, 1996.

[5] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*. Springer, 2006.

[6] Saugata Basu and Marie-Françoise Roy. Bounding the radii of balls meeting every connected component of semi-algebraic sets. *Journal of Symbolic Computation*, 45(12):1270 – 1279, 2010.

[7] P. Bell, V. Halava, T. Harju, J. Karhumäki, and I. Potapov. Matrix equations and Hilbert's Tenth Problem. *IJAC*, 18(8):1231–1241, 2008.

[8] P. Bell, M. Hirvensalo, and I. Potapov. The identity problem for matrix semigroups in $\mathrm{SL}(2, \mathbb{Z})$ is NP-complete. In *SODA*, 2017.

[9] Paul Bell and Igor Potapov. On undecidability bounds for matrix decision problems. *Theoretical Computer Science*, 391(1):3–13, 2008. Combinatorics, Automata and Number Theory.

[10] Paul C. Bell, Igor Potapov, and Pavel Semukhin. On the mortality problem: From multiplicative matrix equations to linear recurrence sequences and beyond. *Information and Computation*, 281:104736, 2021.

[11] A. M. Ben-Amram, S. Genaim, and A. N. Masud. On the termination of integer loops. *ACM Trans. Program. Lang. Syst.*, 34(4), 2012.

[12] J. Berstel. Sur les pôles et le quotient de hadamard de séries n-rationnelles. *C. R. Acad. Sci. Paris Sér. A-B*, 272:A1079-A1081, 1971.

[13] Jean Berstel and Christophe Reutenauer. *Noncommutative rational series with applications*. Cambridge University Press, 2011.

[14] Yuri Bilu, Florian Luca, Joris Nieuwveld, Joël Ouaknine, David Purser, and James Worrell. Skolem Meets Schanuel. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, 2022.

[15] Yuval Bistritz and Alexander Lifshitz. Bounds for resultants of univariate and bivariate polynomials. *Linear Algebra and its Applications*, 432(8):1995 – 2005, 2010.

[16] V. Blondel, E. Jeandel, P. Koiran, and N. Portier. Decidable and undecidable problems about quantum automata. *SIAM J. Comput.*, 34(6):1464–1473, 2005.

[17] M. Braverman. Termination of integer linear programs. In *Proc. Intern. Conf. on Computer Aided Verification (CAV)*, volume 4144 of *LNCS*. Springer, 2006.

[18] J.-Y. Cai. Computing Jordan normal forms exactly for commuting matrices in polynomial time. *Int. J. Found. Comput. Sci.*, 5(3/4):293–302, 1994.

[19] J.-Y. Cai, R. J. Lipton, and Y. Zalcstein. The complexity of the A B C problem. *SIAM J. Comput.*, 29(6), 2000.

[20] Ventsislav Chonev, Joël Ouaknine, and James Worrell. On the complexity of the orbit problem. *J. ACM*, 63(3), June 2016.

[21] Ventsislav Chonev, Joël Ouaknine, and James Worrell. On the Skolem Problem for continuous linear dynamical systems. In *ICALP*, pages 100:1–100:13, 2016.

[22] Henri Cohen. *A Course in Computational Algebraic Number Theory.* Springer, 1993.

[23] Joel E. Cohen and Peter H. Sellers. Sets of nonnegative matrices with positive inhomogeneous products. *Linear Algebra and its Applications*, 47:185–192, 1982.

[24] T. Colcombet, J. Ouaknine, P. Semukhin, and J. Worrell. On Reachability Problems for Low-Dimensional Matrix Semigroups. In *ICALP*, 2019.

[25] George E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decompostion. In *Automata Theory and Formal Languages*, pages 134–183. Springer Berlin Heidelberg, 1975.

[26] C. Conley. *Isolated invariant sets and the Morse index*, volume 25 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1978.

[27] George Dantzig. *Linear programming and extensions*. Princeton University Press, 1963.

[28] Julian D'Costa, Toghrul Karimov, Rupak Majumdar, Joël Ouaknine, Mahmoud Salamati, Sadegh Soudjani, and James Worrell. The Pseudo-Skolem Problem is Decidable. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, 2021.

[29] Julian D'Costa, Toghrul Karimov, Rupak Majumdar, Joël Ouaknine, Mahmoud Salamati, and James Worrell. The pseudo-reachability problem for diagonalisable linear dynamical systems. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, 2022.

[30] Julian D'Costa, Engel Lefaucheux, Eike Neumann, Joël Ouaknine, and James Worrell. On the Complexity of the Escape Problem for Linear Dynamical Systems over Compact Semialgebraic Sets. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, 2021.

[31] Julian D'Costa, Engel Lefaucheux, Eike Neumann, Joël Ouaknine, and James Worrell. Bounding the Escape Time of a Linear Dynamical System over a Compact Semialgebraic Set. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, 2022.

[32] Julian D'Costa, Engel Lefaucheux, Eike Neumann, Joël Ouaknine, and James Worrell. Bounding the escape time of a linear dynamical system over a compact semialgebraic set, 2022. arXiv:2207.01550.

[33] Julian D'Costa, Engel Lefaucheux, Joël Ouaknine, and James Worrell. How fast can you escape a compact polytope? In *Symposium on Theoretical Aspects of Computer Science (STACS)*, 2020.

[34] Julian D'Costa, Joël Ouaknine, and James Worrell. Nonnegativity problems for matrix semigroups. In *Symposium on Theoretical Aspects of Computer Science (STACS)*, 2024.

[35] Papri Dey, Ravi Kannan, Nick Ryder, and Nikhil Srivastava. Bit Complexity of Jordan Normal Form and Polynomial Spectral Factorization. In *ITCS*, 2023.

[36] R. Dong. On the identity problem for unitriangular matrices of dimension four. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, volume 241, pages 43:1–43:14, 2022.

[37] R. Dong. The Identity Problem in $\mathbb{Z} \wr \mathbb{Z}$ Is Decidable. In *ICALP*, volume 261, pages 124:1–124:20, 2023.

[38] Klaus Dräger. The invariance problem for matrix semigroups. In Bart Jacobs and Christof Löding, editors, *Foundations of Software Science and Computation Structures*, pages 479–492, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[39] G. Everest, A. van der Poorten, I. Shparlinski, and T. Ward. *Recurrence Sequences.* American Mathematical Society, 2003.

[40] Nathanaël Fijalkow, Joël Ouaknine, Amaury Pouly, João Sousa-Pinto, and James Worrell. On the decidability of reachability in linear time-invariant systems. In *HSCC*, page 77–86, 2019.

[41] G. Ge. *Algorithms Related to Multiplicative Representations of Algebraic Numbers.* PhD thesis, U.C. Berkeley, 1993.

[42] Balázs Gerencsér, Vladimir V. Gusev, and Raphaël M. Jungers. Primitive sets of nonnegative matrices and synchronizing automata. *SIAM Journal on Matrix Analysis and Applications*, 39(1):83–98, 2018.

[43] Dima Grigoriev and Nicolai Vorobjov. Solving systems of polynomial inequalities in subexponential time. *J. Symbolic Computation*, 5:37 – 64, 1988.

[44] Emmanuel Hainry. Decidability and Undecidability in Dynamical Systems. Research report, Inria, 2009.

[45] 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.

[46] Vesa Halava, Tero Harju, and Mika Hirvensalo. Undecidability bounds for integer matrices using claus instances. *International Journal of Foundations of Computer Science*, 18(05):931–948, 2007.

[47] Vesa Halava and Mika Hirvensalo. Improved matrix pair undecidability results. *Acta Informatica*, 44(3-4):191–205, 2007.

[48] Godfrey H. Hardy and Edward M. Wright. *An Introduction to the Theory of Numbers.* Oxford University Press, 1999.

[49] M.A. Harrison. Lectures on linear sequential machines. Technical report, DTIC Document, 1969.

[50] Joos Heintz, Marie-Françoise Roy, and Pablo Solernó. Sur la complexité du princie de Tarski-Seidenberg. *Bull. Soc. Math. France*, 118(1):101–126, 1990.

[51] Mehran Hosseini, Joël Ouaknine, and James Worrell. Termination of Linear Loops over the Integers. In *ICALP*, 2019.

[52] Gabriella Jeronimo, Daniel Perrucci, and Elias Tsigaridas. On the minimum of a polynomial function on a basic closed semialgebraic set and applications. *SIAM Journal on Optimization*, 23(1):241–255, 2013.

[53] Ravindran Kannan and Richard J. Lipton. Polynomial-time algorithm for the orbit problem. *J. ACM*, 33(4):808–821, 1986.

[54] Toghrul Karimov, Edon Kelmendi, Joris Nieuwveld, Joël Ouaknine, and James Worrell. The power of positivity. In *2023 38th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, pages 1–11, 2023.

[55] Toghrul Karimov, Edon Kelmendi, Joël Ouaknine, and James Worrell. What's decidable about discrete linear dynamical systems?, 2022. arXiv:2206.11412.

[56] Toghrul Karimov, Edon Kelmendi, Joël Ouaknine, and James Worrell. Multiple reachability in linear dynamical systems, 2024. arXiv:2403.06515.

[57] S.-K. Ko, R. Niskanen, and I. Potapov. On the identity problem for the special linear group and the Heisenberg group. In *ICALP*, pages 132:1–132:15, 2018.

[58] S. Lang. Introduction to transcendental numbers. Addison-Wesley Series in Mathematics. Reading, Mass. etc.: Addison-Wesley Publishing Company. VI, 105 p. (1966)., 1966.

[59] A.K. Lenstra, H.W. Lenstra Jr., and Lászlo Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261:515–534, 1982.

[60] Richard Lipton, Florian Luca, Joris Nieuwveld, Joël Ouaknine, David Purser, and James Worrell. On the skolem problem and the skolem conjecture. In *LICS*, 2022.

[61] Florian Luca, James Maynard, Armand Noubissie, Joël Ouaknine, and James Worrell. Skolem meets Bateman-Horn, 2024.

[62] Angus Macintyre. Turing meets Schanuel. *Annals of Pure and Applied Logic*, 167(10):901–938, 2016. Logic Colloquium 2012.

[63] Angus Macintyre and Alex J. Wilkie. On the decidability of the real exponential field. In Piergiorgio Odifreddi, editor, *Kreiseliana. About and Around Georg Kreisel*, pages 441–467. A K Peters, 1996.

[64] A. Markov. On certain insoluble problems concerning matrices. *Doklady Akad. Nauk SSSR*, 57(6):539–542, June 1947.

[65] Pavel Martyugin. Computational complexity of certain problems related to carefully synchronizing words for partial automata and directing words for nondeterministic automata. *Theor. Comp. Sys.*, 54(2):293–304, February 2014.

[66] D. W. Masser. Linear relations on algebraic groups. In *New Advances in Transcendence Theory*, page 248–262. Camb. Univ. Press, 1988.

[67] Carl D. Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2023.

[68] M. Mignotte. Some useful bounds. In *Computer Algebra*, 1982.

[69] M. Mignotte, T. Shorey, and R. Tijdeman. The distance between terms of an algebraic recurrence sequence. *J. für die reine und angewandte Math.*, 349, 1984.

[70] Kenji Nagasaka and Jau-Shyong Shiue. Asymptotic positiveness of linear recurrence sequences. *Fibonacci Quart*, 28(4):340–346, 1990.

[71] Eike Neumann, Joël Ouaknine, and James Worrell. On ranking function synthesis and termination for polynomial programs. In *CONCUR*, 2020.

[72] D. Noutsos. On perron–frobenius property of matrices having some negative entries. *Linear Algebra and its Applications*, 412(2-3):132–153, 2006.

[73] J. Ouaknine, A. Pouly, J. Sousa-Pinto, and J. Worrell. Solvability of matrix-exponential equations. In *LICS*, pages 798–806, 2016.

[74] J. Ouaknine and J. Worrell. Positivity problems for low-order linear recurrence sequences. In *SODA*, 2014.

[75] J. Ouaknine and J. Worrell. Ultimate Positivity is decidable for simple linear recurrence sequences. In *Proceedings of ICALP'14*, 2014. *CoRR, abs*/1309.1914.

[76] Joël Ouaknine and James Worrell. On the positivity problem for simple linear recurrence sequences. In *Automata, Languages, and Programming: 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part II 41*, pages 318–329. Springer, 2014.

[77] Joël Ouaknine and James Worrell. On linear recurrence sequences and loop termination. *ACM SIGLOG News*, 2015.

[78] V. Pan. Optimal and nearly optimal algorithms for approximating polynomial zeros. *Computers & Mathematics with Applications*, 31(12), 1996.

[79] Christos H. Papadimitriou and Georgios Piliouras. From nash equilibria to chain recurrent sets: An algorithmic solution concept for game theory. *Entropy*, 20(10):782, 2018.

[80] M. S. Paterson. Undecidability in 3 by 3 matrices. *J. of Math. and Physics*, 1970.

[81] V. Yu. Protasov. Classification of \$k\$-primitive sets of matrices. *SIAM Journal on Matrix Analysis and Applications*, 34(3):1174–1188, 2013.

[82] J. Renegar. On the computational complexity and geometry of the first-order theory of the reals. *J. Symb. Comp.*, 13(3):255 – 352, 1992.

[83] D. Richardson. How to recognize zero. *Journal of Symbolic Computation*, 24(6):627–645, 1997.

[84] Siegfried M. Rump. Polynomial minimum root separation. *Mathematics of Computation*, 33(145):327–336, 1979.

[85] Akshay S., Nikhil Balaji, and Nikhil Vyas. Complexity of Restricted Variants of Skolem and Related Problems. In Kim G. Larsen, Hans L. Bodlaender, and Jean-Francois Raskin, editors, *42nd International Symposium on Mathematical Foundations of Computer Science (MFCS 2017)*, volume 83 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 78:1–78:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[86] Marcus Schaefer and Daniel Stefankovic. Fixed points, nash equilibria, and the existential theory of the reals. *Theory Comput. Syst.*, 60(2):172–193, 2017.

[87] Hans Schneider. Wielandt's proof of the exponent inequality for primitive nonnegative matrices. *Linear Algebra and its Applications*, 353(1):5–10, 2002.

[88] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[89] Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer-Verlag, Berlin, Heidelberg, 1998.

[90] T. Tao. *Structure and randomness: pages from year one of a mathematical blog*. American Mathematical Society, 2008.

[91] A. Tarski. *A Decision Method for Elementary Algebra and Geometry*. University of California Press, 1951.

[92] A. Tiwari. Termination of linear programs. In *Proc. Intern. Conf. on Comp. Aided Verif. (CAV)*, volume 3114 of *LNCS*. Springer, 2004.

[93] Lou van den Dries. Remarks on Tarski's problem concerning (R, +, *, exp). In G. Lolli, G. Longo, and A. Marcja, editors, *Logic Colloquium '82*, Studies in Logic and the Foundations of Mathematics. North-Holland, Amsterdam, 1984.

[94] N. K. Vereshchagin. The problem of appearance of a zero in a linear recurrence sequence (in Russian). *Mat. Zametki*, 38(2), 1985.

[95] Ernest Borisovich Vinberg, Arkadij L. Onishchik, and Dimitry Leites. *Lie groups and algebraic groups*. Springer Verlag, 1990.

[96] Nicolai Vorobjov. Bounds of real roots of a system of algebraic equations. *Zap. Nauchn. Sem. LOMI*, 137:7 – 19, 1984. (in Russian).

[97] A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094, 1996.

[98] A. J. Wilkie. Schanuel's conjecture and the decidability of the real exponential field. In *Algebraic Model Theory*, pages 223–230. Springer Netherlands, Dordrecht, 1997.

[99] Yaokun Wu and Yinfeng Zhu. Primitivity and hurwitz primitivity of nonnegative matrix tuples: A unified approach. *SIAM Journal on Matrix Analysis and Applications*, 44(1):196–211, 2023.