SEMIDEFINITE REPRESENTATIONS IN SEMIALGEBRAIC OPTIMIZATION AND DYNAMICS-ORIENTED LEARNING

Bachir El Khadir

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Operations Research and Financial Engineering
Adviser: Amir Ali Ahmadi

September 2020

Abstract

We study the power and limitations of semidefinite programming for representing semialgebraic functions that satisfy various nonnegativity constraints. In Part I, this is done in the context of semialgebraic optimization, and in Part II, in the context of dynamical systems.

We start Part I by introducing the framework of time-varying semidefinite programs (TV-SDPs). TV-SDPs, which can be seen as a broad generalization of the notion of continuous linear programs introduced by Bellman in 1953, are semidefinite programs (SDPs) whose data and solutions are allowed to vary with time. For any TV-SDP that satisfies some mild assumptions, we show that polynomial functions of time are arbitrarily close to being optimal, and that the best polynomial solution of a given degree can be found by solving an SDP of tractable size. We also show that certificates of near optimality can be computed via SDP.

We then turn our attention to studying the interplay between two notions that are known to make semialgebraic optimization problems more tractable: the algebraic notion of a sum of squares decomposition and the geometric notion of convexity. In 2009, Blekherman showed that for high enough number of variables, there must be convex forms of degree 4 that are not sums of squares. Remarkably, no examples are known to date. We show that all degree-4 convex forms are sums of squares when the number of variables is less than or equal to 4. A main ingredient of the proof is the derivation of certain "generalized Cauchy-Schwarz inequalities," which could be of independent interest.

We start Part II by studying SDP-based approaches for certifying properties of dynamical systems via semialgebraic Lyapunov functions. We give the first example of a globally asymptotically stable polynomial vector field with rational coefficients that does not have a polynomial Lyapunov function, even locally. By contrast, we show that an asymptotically stable *homogeneous* vector field admits a Lyapunov function that is a polynomial divided by a power of the 2-norm of the state variable. We further show that this Lyapunov function can be found via semidefinite programming.

We finally propose an SDP-based approach for the problem of learning a dynamical system from noisy observations of a few trajectories and subject to *side information* (i.e., any knowledge that we may have about the dynamical system besides trajectory data). We identify six types of side information that arise naturally in applications and show that they can be imposed on polynomial vector fields via semidefinite programming. We study how well polynomial vector fields can approximate continuously-differentiable ones while satisfying side information. We end by showing the applicability of our framework to imitation learning problems in robotics.

Acknowledgements

I am forever grateful to my advisor Amirali Ahmadi for his guidance, constant support, encouragement, and generosity. He never fails to make time for his students. I fondly recall the numerous meetings we devoted to exploring new research ideas, polishing the draft of a paper, or designing the plan for a conference talk. Thank you for making my PhD years memorable. I would also like to thank my thesis reader and committee members: Bartolomeo Stellato, Ben Recht, Bob Vanderbei, and Ramon Van Handel.

I owe a lot to Gökçe for her unwavering support to me. I would also like to thank my research group: Cemil, Georgina, Jeff and all my friends at ORFE: Aldo, Cong, Elahe, Emmanuel, Jacob, Kaizheng, Kobe, Levon, Mark, Nongchao, Pierre-Yves, Samy, Thomas, Wenyan, Yair, Zach, Zhuoran, and Zongxi.

Special thanks to Vikas Sindhwani for hosting me at Google for the summer of 2018, where I had the chance to collaborate with the wonderful Robotics team. I am also very grateful to Jean-Bernard Lasserre for hosting me for a week in LAAS in Toulouse, where we spent hours on the board trying to make sense of some intricate control problems.

اتقدم بجزيل الشكر و الامتنان لأمي العزيزة كنزة و لأختاي ايمان و أميمة من أجل دعمهم و مساندتهم لي طوال سنوات الدراسة.

To my father Abderrahmane.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The content of this thesis traces back to the work of two prominent mathematical figures of the 19th century: the work of David Hilbert on sum of squares (sos) representations of nonnegative polynomials, and the work of Aleksandr Lyapunov on dynamical systems analysis. While initially considered remote from practical applications, their work has been recently brought to center stage in the optimization and control communities thanks to some recent theoretical and algorithmic developments in the area of semidefinite programming (SDP)[1]. These developments have in turn enabled a wealth of new applications in the sciences and engineering.

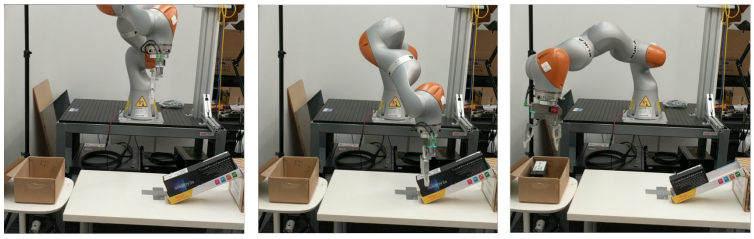
Our broad goal in this thesis is to shed some light on the fascinating connections between the worlds of SDP, sos representations of polynomials, and dynamical systems theory. More concretely, we study the power and limitations of SDP-based approaches to (i) solving semialgebraic optimization problems , (ii) verifying stability properties of dynamical systems, and (iii) learning dynamical systems from data. Before we give a formal definition of semidefinite programming, let us preview in Figure 1.1 some of its applications to semialgebraic optimization and dynamical systems theory that are studied in this thesis. In Figure 1.1a, a robotic arm has learned, by solving an SDP, to autonomously execute a demonstrated task in an environment filled with obstacles. Figure 1.1b depicts an instance of a time-varying maximum flow problem together with a near-optimal flow that was computed via SDP. Figure 1.1c shows a streamplot of a polynomial vector field for which a natural SDP-based approach fails to prove stability.

1.1 Semidefinite Programming

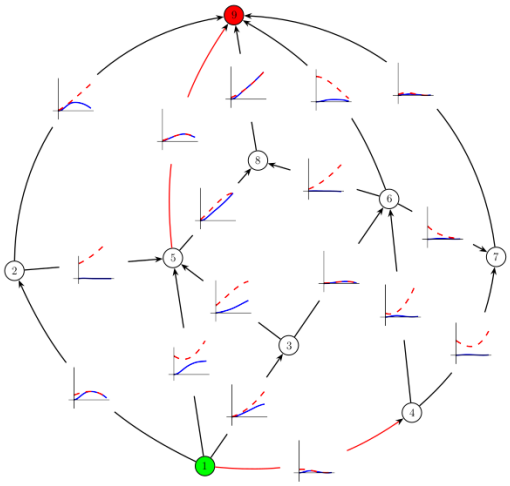
We denote by S_n the space of $n \times n$ symmetric matrices equipped with the inner product

$$\langle A, B \rangle := \sum_{i,j=1}^{n} A_{ij} B_{ij} \quad \forall A, B \in S_n,$$

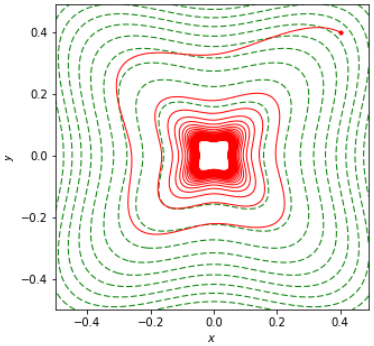
[1]When it is clear for the context, we use the acronym SDP to denote both "semidefinite programming" and "semidefinite program".



(a) A robotic arm picking objects and placing them in a bin. The robot learns how to accomplish the task from a single human demonstration by solving an SDP. See chapter 7 for more details.



(b) An instance of a time-varying maxflow problem studied in Chapter 2. On every edge, we plot in red the edge capacities as a function of time, and in blue the optimal flow among polynomial functions of time of degree 10. This flow was computed by solving an SDP.



(c) Streamplot of a polynomial vector field that is asymptotically stable but does not admit a polynomial Lyapunov function even locally. This is our main construction in Chapter 5.

Figure 1.1: Preview of some applications of semidefinite programming that are analyzed in this thesis. Figures 1.1a and 1.1b are instances where semidefinite programming produces good solutions, empirically or theoretically. Figure 1.1c shows a possible limitation of SDP-based methods in dynamical systems theory.

and the partial ordering $\succeq$ given by

$$\left[A \succeq B \iff x^T(A-B)x \geq 0 \quad \forall x \in \mathbb{R}^n\right] \quad \forall A, B \in S_n.$$

Semidefinite programming is the task of minimizing a linear function $\langle C, X\rangle$ over the set of positive semidefinite matrices

$$S_n^+ := \{X \in S_n \mid X \succeq 0\}$$

subject to affine constraints

$$\langle A_i, X\rangle = b_i, \quad i = 1, \ldots, m.$$

Here, C, A_1, …, A_m are symmetric $n \times n$ matrices, and b is an $n \times 1$ vector. Semidefinite programming contains as special case some of the most common classes of convex optimization problems, including linear programs, convex quadratic programs, and second-order cone programs.

Among the factors that have contributed to the popularization of SDPs are some recent algorithmic developments and powerful theoretical results. On the algorithmic side, interior point methods, that were initially developed for linear programs, were adapted for the setting of semidefinite programming; they allow one to solve SDPs to arbitrary accuracy in polynomial time [202]. On the theoretical side, semidefinite programming enjoys a rich duality theory. Indeed, every SDP

$$\begin{array}{ll} \min\limits_{X \in S_n^+} & \langle C, X\rangle \\ \text{subject to} & \langle A_i, X\rangle = b_i, \quad i = 1, \ldots, m, \end{array} \tag{1.1}$$

comes with a *dual* program

$$\begin{array}{ll} \max\limits_{x \in \mathbb{R}^m} & \langle b, x\rangle \\ \text{subject to} & C - \sum_{i=1}^m x_i A_i \succeq 0, \end{array} \tag{1.2}$$

that is also an SDP. The objective value attained by any feasible solution to the dual program (1.2) is always a lower bound on the optimal value of the *primal* program (1.1). Under mild assumptions, there is no duality gap, i.e., the optimal values of (1.1) and (1.2) match. The ability to quantify how far a feasible solution is from global optimality is a desirable property in many applications. See [202] for a survey on the theoretical and algorithmic aspects of semidefinite programming.

Throughout the years, semidefinite programming has been successfully applied in diverse areas such as combinatorial optimization [82], control theory[49], computational geometry [42], and statistics [34]. In fact, there are several mathematical problems where the state-of-the-art methods are based on semidefinite programming [82, 51]. This observation calls for a more thorough study of the power of SDP-based methods.

Most of the content of this thesis connects semidefinite programming to two application areas: semialgebraic optimization and dynamical systems theory. The next two

sections of this introduction will make this connection clearer. In Chapter 2, however, we extend some theoretical and algorithmic aspects of semidefinite programming itself by considering a setting where the data of a semidefinite program vary with time. Among the questions studied in Chapter 2 are: (i) How can one compute upper bounds and lower bounds on the optimal value of an SDP with time-varying data? and (ii) Under what conditions can the difference between the two bounds be made arbitrarily small?

1.2 SDP-Based Methods in Semialgebraic Optimization

By and large, the most common functions used for modeling purposes in optimization are the linear functions. This is mostly due to their simplicity, and the availability of software dedicated to linear models. Stepping into the nonlinear world, one encounters *polynomial* functions. One of the most basic questions that an optimizer can ask about a polynomial p is whether it is nonnegative, i.e., whether

$$p(x) \geq 0 \quad \forall x \in \mathbb{R}^n. \tag{1.3}$$

An efficient way to check nonnegativity of polynomial functions naturally leads to an efficient method to minimize (or maximize) polynomial functions as well. Indeed, consider the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} \quad p(x). \tag{1.4}$$

The optimal value of (1.4) is equal to the optimal value of the following maximization problem (over the scalar γ):

$$\max_{\gamma \in \mathbb{R}} \quad \gamma \text{ s.t. } p - \gamma \text{ is nonnegative.} \tag{1.5}$$

Given an efficient way of testing nonnegativity, one can, for example, perform bisection on γ to find the optimal value of (1.4) (whenever finite lower and upper bounds on the optimal value are available).

1.2.1 Sum of squares programming

Unfortunately, testing whether a polynomial is nonnegative is an NP-hard problem already for degree-4 polynomials [144]. In 1888 [97], for reasons that were different from the computational considerations mentioned here, Hilbert studied the question of whether every nonnegative polynomial p could be represented as a sum of squares of other polynomials, i.e., whether

$$\text{there exist polynomials } h_1, \ldots, h_r \text{ s.t. } p = \sum_{i=1}^{r} h_i^2. \tag{1.6}$$

He showed that the answer is negative, i.e., that there exist nonnegative polynomials that are not sums of squares (sos). Still, the representation (1.6) is appealing from an optimization point of view. Indeed, it acts as an algebraic *certificate* of nonnegativity of the polynomial p: given polynomials $h_1, \ldots, h_r$, one can check that equality (1.6) holds by expanding the sum on its right hand side and comparing the coefficients of the resulting polynomial to those of p. By contrast, using the definition of nonnegativity in (1.3) directly requires the evaluation of the polynomial p at infinitely many points. Moreover, it can be shown [156] that a degree-d polynomial p admits an sos representation as in (1.6) if and only if there exists a positive semidefinite matrix Q that satisfies the identity

$$p(x) = z(x)^T Q z(x), \tag{1.7}$$

where $z(x) := (1, x_1, x_2, x_1x_2, \ldots, x_n^{\frac{d}{2}})^T$ denotes the vector of all monomials in x of degree less than or equal to $d/2$. Searching for such a matrix Q is an SDP, which can be solved efficiently. This remains true even if some of the coefficients of p are unknown. The technique of replacing nonnegativity constraints with an SDP that searches for an appropriate sos decompositions, such as the one in (1.6), is called *sum of squares programming*.

Sos programming can be extended to the *constrained* setting. Suppose we are interested in testing nonnegativity of a polynomial p over a *closed basic semialgebraic* set K (i.e., a set that is defined by finitely many polynomial inequalities $g_1 \geq 0, \ldots, g_m \geq 0$). Then, a decomposition of p as

$$p(x) = \sigma_0(x) + \sum_{i=1}^{m} \sigma_i(x) g_i(x), \tag{1.8}$$

where the polynomials σ_i are sos, is a certificate of nonnegativity of p over the set K. In a similar fashion to the unconstrained case, the search for the sos polynomials σ_i can be cast as an SDP whenever a bound on their degrees is imposed.

Another important extension of sos programming deals with *semialgebraic* functions and leads to new techniques to tackle *semialgebraic* optimization problems. We give a definition of semialgebraic functions and semialgebraic optimization below.

Definition 1.2.1. *A set is closed semialgebraic if it is a finite union of closed basic semialgebraic sets. A function is called semialgebraic if its epigraph is a closed semialgebraic set. An optimization problem is called semialgebraic if its objective function is semialgebraic, and its feasible set is closed semialgebraic.*

Examples of semialgebraic functions include polynomials of course, but also *rational* functions (i.e., ratio of polynomials), the square root function, and the absolute value function. Extensions of the sum of squares approach often allow for an automated search for algebraic certificates of nonnegativity of semialgebraic functions over semialgebraic sets via semidefinite programming.

1.2.2 Analyzing sum of squares relaxations

The sos programming approach described in the previous subsection usually trades off "exactness for efficiency". For example, for the problem (1.5) of unconstrained minimization of a given polynomial p, replacing the nonnegativity constraint with an sos constraint leads to a lower bound on the minimum value of p. This lower bound can be computed efficiently via SDP, but is not sharp in general. A lot of effort goes into studying special structural properties of the problem under which sos programming leads to exact solutions or, at least, to solutions that are good enough for the problem at hand. For instance, one of the main results of Chapter 2 relies on the existence of (small) sos certificates of nonnegativity for *univariate* polynomial matrices. Another example where we leverage special properties of the problem at hand appears in Chapter 4. We show there that for *homogeneous* polynomial vector fields, asymptotic stability is equivalent to the existence of a rational Lyapunov function, and that Lyapunov inequalities on both this rational function and its time derivative have sos certificates. Finally, in Chapters 6 and 7, we analyze the power of sos programming for the task of fitting a polynomial vector field to trajectory data while imposing certain properties on the trajectories of this vector field.

1.2.3 When sos representations meet convex analysis

When a semialgebraic optimization problem enjoys *convexity* properties, one can use either sos programming or gradient descent-based methods from nonlinear optimization to tackle the problem. Each one of these two approaches leverages a different property of the optimization problem, and it is not at all obvious how to design an algorithm that can benefit from the power of these two approaches at the same time. Indeed, consider problem (1.4) for example, and assume that the polynomial p that appears as the objective function is convex. On the one hand, gradient descent methods with small enough step size are guaranteed to converge to a global minimizer of p, but fail to take advantage of the fact that the objective function has a special algebraic structure. On the other hand, sos programming makes no assumption about convexity of the polynomial p, and as a result, does not explicitly exploit this property. This observation motivates our study in Chapter 3, where we focus on the intriguing interplay between the geometric notion of convexity and the algebraic notion of a sum of squares decomposition.

1.3 Power and Limitations of SDP in Dynamical Systems Theory

An area that has historically generated a lot of interest in SDP is stability analysis of equilibrium points of dynamical systems. Consider a dynamical system

$$\dot{x} = f(x) \tag{1.9}$$

with $f : \mathbb{R}^n \to \mathbb{R}^n$, and an equilibrium point x_e (i.e., a point $x_e \in \mathbb{R}^n$ with $f(x_e) = 0$). Roughly speaking, the equilibrium point x_e is *stable in the sense of Lyapunov* if trajectories that start close enough to that point remain close enough forever. If, in addition to being stable in the sense of Lyapunov, trajectories that start near x_e (resp. all trajectories in $\mathbb{R}^n$) converge to x_e, then x_e is said to be *locally asymptotically stable* (resp. *globally asymptotically stable*).

At first glance, it may seem that proving any stability property of a dynamical system requires solving the differential equation in (1.9) analytically. This approach, however, is impractical for most nonlinear dynamical systems. The key insight behind Lyapunov theory is to turn the question of testing stability into a search for an energy-like function V (called a *Lyapunov function*) that decreases along trajectories. For instance, under mild assumptions on the dynamical system $\dot{x} = f(x)$, global asymptotic stability of an equilibrium point x_e is equivalent [112] to the existence of a radially unbounded function $V : \mathbb{R}^n \to \mathbb{R}^n$ that satisfies the following Lyapunov inequalities:

$$V(x) > 0 \quad \text{and} \quad -\dot{V}(x) := -\langle \nabla V(x), f(x) \rangle > 0 \quad \text{for all } x \neq x_e. \tag{1.10}$$

In practice, one restricts the search for Lyapunov functions to a finite dimensional family of functions, and uses optimization techniques to find one that satisfies the requirements in (1.10). Let us illustrate this process for the particularly nice setting of *linear* dynamical systems. It is a classical result that the origin of a linear dynamical system $\dot{x} = Ax$ (where A is a square matrix) is asymptotically stable[2] if and only if this system admits a *quadratic* Lyapunov function V, say

$$V(x) = x^T Q x, \quad \text{where } Q \in S_n. \tag{1.11}$$

A function of the form (1.11) is a valid Lyapunov function if and only if

$$Q \succ 0 \quad \text{and} \quad -QA - A^T Q \succ 0. \tag{1.12}$$

By multiplying the matrix Q by an appropriate positive constant, we can replace the constraints in (1.12) with the equivalent constraints

$$Q \succeq I \quad \text{and} \quad -QA - A^T Q \succeq I,$$

where I denotes the identity matrix. Checking whether a linear dynamical system is asymptotically stable can therefore be done efficiently via SDP.

In this thesis, we are mostly interested in the setting where the vector field f in (1.9) is a polynomial function. In that case, it is natural to combine tools from Lyapunov theory and semialgebraic optimization to certify stability properties of an equilibrium point of f. Indeed, a common approach in the literature is to parameterize

[2]For homogeneous dynamical systems (in particular, for linear dynamical systems), global asymptotic stability is equivalent to local asymptotic stability. Therefore, we refer to an equilibrium point of a homogeneous dynamical system that satisfies one (and thus both) of the latter properties simply as asymptotically stable.

the candidate function V itself as a polynomial of a fixed degree. The Lyapunov inequalities arising from (1.10) lead then to the search for a polynomial function V that satisfies the constraints[3]

$$V \text{ sos and } -\dot{V} \text{ sos.}$$

As we have seen, this search can be carried out via SDP.

At this point, one might wonder whether stability of polynomial vector fields can always be established with the help of polynomial Lyapunov functions. The answer is known to be negative when the notion of stability of interest is global asymptotic stability [12], and when the notion of stability of interest is local asymptotic stability and the coefficients of the polynomial vector field are allowed to be irrational [29]. We construct in Chapter 5 the first example of a polynomial vector field with rational coefficients that is locally asymptotically stable but does not admit an analytic Lyapunov functions, let alone a polynomial one.

Considerable effort is devoted to generalizing the search for Lyapunov functions beyond the class of polynomial functions, and to finding suitable assumptions on the vector field f under which Lyapunov functions can be efficiently computed. In Chapter 4, we show that an equilibrium point of a homogeneous polynomial vector field is asymptotically stable if and only if there exists a corresponding rational Lyapunov function (i.e., ratio of two polynomials). We further show that the Lyapunov inequalities on both the rational function and its time derivative have sos certificates. Hence such a Lyapunov function can be found by semidefinite programming.

Interestingly, the same tools that help analyze dynamical systems can also be applied for learning these dynamical systems from data. We focus in Chapter 6 on the efficacy of sos programming for the problem of learning a vector field that fits sample trajectories and satisfies a concrete collection of what we call "side information" constraints. We end by showing the applicability of our learning framework for imitation learning problems in robotics in Chapter 7.

1.4 Outline

The outline of this thesis is as follows:

Part I: On the Interface of Semidefinite Programming and Semialgebraic Optimization

The first part of this thesis studies the interplay between semidefinite programming and semialgebraic optimization. In Chapter 2, we study time-varying semidefinite programs, which are SDPs whose data (and solutions) depend on time. Our focus is on the setting where the data vary polynomially with time. In Chapter 3, we study the relationship between two notions that make semialgebraic optimization more tractable: the geometric notion of convexity, and the algebraic notion of a sum of squares decomposition.

[3] In this introduction, we ignore the subtleties arising from the distinction between strict and non strict Lyapunov inequalities.

Part II: Semidefinite Programming for Analyzing and Learning Dynamical Systems

The second part of this thesis focuses on the power and limitations of SDP-based approaches to analyzing dynamical systems and learning them from data. In Chapters 4 and 5, we focus on stability of dynamical systems. We show both positive and negative results concerning the existence of semialgebraic Lyapunov functions. In Chapters 6 and 7, we study how semialgebraic optimization tools can be used in the context of learning dynamical systems from data.

1.5 Related Publications

The publications associated with this thesis are as follows:

- Amir Ali Ahmadi and Bachir El Khadir. Time-Varying Semidefinite Programs. Accepted to Mathematics of Operations Research, 2020.
- Bachir El Khadir. On Sum of Squares Representation of Convex Forms and Generalized Cauchy-Schwarz Inequalities. SIAM Journal on Applied Algebra and Geometry, 4(2), 377–400, 2020.
- Amir Ali Ahmadi and Bachir El Khadir. On Algebraic Proofs of Stability for Homogeneous Vector Fields. IEEE Transactions on Automatic Control 65 (1), 325-332, 2019.
- Amir Ali Ahmadi and Bachir El Khadir. A Globally Asymptotically Stable Polynomial Vector Field with Rational Coefficients and no Local Polynomial Lyapunov Function. Systems & Control Letters 121, 50-53, 2018.
- Amir Ali Ahmadi and Bachir El Khadir. Learning Dynamical Systems with Side Information. In the Proceedings of Machine Learning Research vol 120:1–10, 2020. Full version of the paper available at arXiv:2008.10135.
- Bachir El Khadir, Jack Varley, and Vikas Sindhwani. Teleoperator Imitation with Continuous-time Safety. In the Proceedings of the Robotics: Science and Systems (RSS), 2019.

Part I

On the Interface of Semidefinite Programming and Semialgebraic Optimization

Chapter 2

Time-Varying Semidefinite Programs

2.1 Introduction

We study semidefinite programs (SDPs) whose feasible set and objective function depend on time. More specifically, a *time-varying semidefinite program* (TV-SDP) is an optimization problem of the form

$$\begin{array}{ll} \sup\limits_{x \in \mathbf{L^n}} & \int_0^1 \langle c(t), x(t) \rangle dt \\ \text{subject to} & Fx(t) \succeq 0 \qquad \forall t \in [0,1] \text{ a.e..} \end{array} \tag{2.1}$$

Here, the operator $F : \mathbf{L^n} \to \mathbf{S^m}$ is defined as

$$Fx(t) := A_0(t) + \sum_{i=1}^{n} x_i(t) A_i(t) + \sum_{i=1}^{n} \int_0^t x_i(s) D_i(t,s) ds, \tag{2.2}$$

where

$$\mathbf{L^n} := \{x : [0,1] \to \mathbb{R}^n \mid x \text{ measurable and } \sup_{t \in [0,1], i=1,\ldots,n} |x_i(t)| < \infty\},$$

and

$$\mathbf{S^m} := \{X : [0,1] \to \mathbb{R}^{m \times m} \mid X(t) \text{ is symmetric } \ \forall t \in [0,1] \text{ and } \sup_{t \in [0,1], i,j=1,\ldots,m} |X_{ij}(t)| < \infty\}.$$

The data to the problem consist of $c \in \mathbf{L^n}$, $A_i \in \mathbf{S^m}$ for $i \in \{0, \ldots, n\}$, and D_i for $i \in \{1, \ldots, n\}$, which satisfies the requirement that $D_i(t, .)$ be a measurable function in $\mathbf{S^m}$ for all $t \in [0,1]$ and that $\sup_{t,s \in [0,1]} \|D_i(t,s)\| < \infty$, where $\|.\|$ is any matrix norm. For a symmetric matrix M, we write $M \succeq 0$ to denote that M is positive semidefinite, i.e., has nonnegative eigenvalues. The abbreviation *a.e.* indicates that the matrix inequality in (2.1) should hold "almost everywhere"; i.e., for

every $t \in [0,1] \setminus N$, where N is some set of measure zero with respect to the Lebesgue measure.

For an interval $I \subseteq \mathbb{R}$, we define the set

$$\mathbf{S^{m+}}(I) := \{X \in \mathbf{S^m} \mid X(t) \succeq 0 \quad \forall t \in I \text{ a.e.}\}.$$

With this notation, a *feasible solution* to the TV-SDP in (2.1) is a function $x \in \mathbf{L^n}$ that satisfies the constraint

$$Fx \in \mathbf{S^{m+}}([0,1]), \tag{2.3}$$

and the *feasible set* of the TV-SDP is the set $\mathcal{F} := \{x \in \mathbf{L^n} \mid Fx \in \mathbf{S^{m+}}([0,1])\}$. The choice of the interval $[0,1]$ is of course made for convenience. Without loss of generality, we can reduce any bounded interval $[a,b]$, with $a < b$, to the interval $[0,1]$ by performing the change of variable $t' = \frac{t-a}{b-a}$.

We equip $\mathbf{L^n}$ and $\mathbf{S^m}$ respectively with the inner products $\langle \cdot, \cdot \rangle_{\mathbf{L^n}}$ and $\langle \cdot, \cdot \rangle_{\mathbf{S^m}}$ defined as

$$\langle x, y \rangle_{\mathbf{L^n}} := \int_0^1 \langle x(t), y(t) \rangle \, dt = \sum_{i=1}^n \int_0^1 x_i(t) y_i(t) \, dt,$$

and

$$\langle X, Y \rangle_{\mathbf{S^m}} := \int_0^1 \langle X(t), Y(t) \rangle \, dt = \int_0^1 \operatorname{Tr}(X(t)Y(t)) \, dt,$$

where $\operatorname{Tr}(A)$ stands for the trace of a matrix A. Using the notation for the first inner product above, the TV-SDP in (2.1) can be written more compactly as

$$\begin{array}{ll} \sup\limits_{x \in \mathbf{L^n}} & \langle c, x \rangle_{\mathbf{L^n}} \\ \text{subject to} & Fx \in \mathbf{S^{m+}}([0,1]). \end{array}$$

The terms $\int_0^t x_i(s) D_i(t,s) ds$ in (2.2) are called *kernel terms* and broaden the class of problems that can be modelled as a TV-SDP. The special case where the terms $D_i(t,s)$ are identically zero is already interesting and presents an infinite sequence of SDPs indexed by time $t \in [0,1]$. While these SDPs are in principle independent of each other, basic strategies such as sampling t and solving a finite number of independent SDPs generally fail to provide a solution to the TV-SDP. This is because candidate functions obtained from simple interpolation schemes can violate feasibility in between sample points. When the terms $D_i(t,s)$ are not zero, the value that a solution takes at a given time affects the range of values that it can take at other times. When the terms $D_i(t,s)$ are constant functions of t and s for instance, the TV-SDP in (2.1) is already powerful enough to express linear constraints involving the function x and its derivatives and/or integrals of any order. For example, to impose a constraint on $x'(t)$, one can introduce a new decision variable $y \in \mathbf{L^n}$ which is related to x via the linear constraint $x(t) - \int_0^t y(s) \, \mathrm{d}s = 0$.

In this chapter, we consider the data $c, A_0, \ldots, A_n, D_1, \ldots, D_n$ of the TV-SDP in (2.1) to belong to the class of *polynomial* functions. Our interest in this setting stems from two reasons. On the one hand, the set of polynomial functions is dense in the set of continuous functions on $[0,1]$ and hence powerful enough for modeling purposes.

On the other hand, polynomials can be finitely parameterized (in the monomial basis for instance) and are very suitable for algorithmic operations.

Even when the input data to a TV-SDP is polynomial, there is no reason to expect its optimal solution to be a polynomial or even a continuous function. Nevertheless, we concern ourselves in this chapter with finding feasible polynomial solutions to a TV-SDP (which naturally provide lower bounds on its optimal value). Our motivation for making this choice is twofold. First, solutions that are smooth are often preferred in practice. Consider for example the problem of scheduling generation of electric power when daily user consumption varies with time, or that of finding a time-varying controller for a robotic arm that serves some routine task in a production line. In such scenarios, smoothness of the solution is important for avoiding deterioration of the hardware or guaranteeing safety of the workplace. Continuity of the solution is even more essential as physical implementation of a discontinuous solution is not viable.

Our second motivation for studying polynomial solutions is algorithmic. We will show (cf. Section 2.3.2) that optimal polynomial solutions of a given degree to a TV-SDP with polynomial data can be found by solving a (non time-varying) SDP of tractable size.

These observations call for a better understanding of the power of polynomial solutions as their degree increases, or a methodology that can bound their gap to optimality when their degree is capped. These considerations are the subjects of Section 2.3.1 and Section 2.4 respectively.

As an illustration of a TV-SDP with polynomially time-varying data and a preview of our solution technique, consider problem (2.1) with $n = 2$ and data

$$A_0(t) = \begin{pmatrix} (1-\frac{8}{5}t)^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, A_1(t) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, A_2(t) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

$$D_1(t,s) = D_2(t,s) = 0, c(t) = \begin{pmatrix} 9t^2 - 9t + 1 \\ 23t^3 - 34t^2 + 12t \end{pmatrix}.$$

As the kernel terms $D_i(t,s)$ are identically zero here, an optimal solution to this TV-SDP is a function $x^{\text{opt}} \in \mathbf{L^2}$ such that for all t in $[0,1]$ (except possibly on a set of measure zero), $x^{\text{opt}}(t)$ is a maximizer of $\langle c(t), x\rangle$ under the constraints $A_0(t) + x_1A_1(t) + x_2A_2(t) \succeq 0$. In Figure 2.1, the dotted red line represents the optimal polynomial solution $x^{\text{poly},20}(t)$ of degree 20. The feasible set

$$\mathcal{F}_t := \{x \in \mathbb{R}^2 \mid A_0(t) + x_1A_1(t) + x_2A_2(t) \succeq 0\}$$

for some sample times t is delimited by blue lines. The objective function $c(t)$ is represented by a black arrow, which also moves in time. The feasible solution $x^{\text{poly},20}(t)$ achieves an objective value of 0.89. By solving an inexpensive dual problem (with $d = 10$ in problem (2.15) of Section 2.4), we can conclude that the optimal value of the

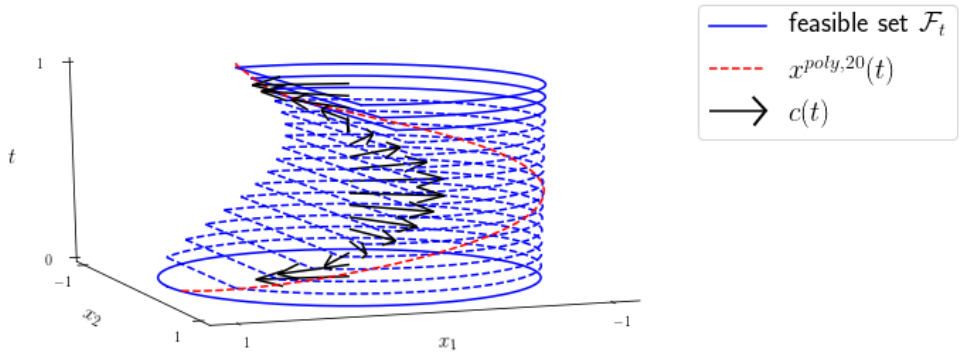


Figure 2.1: An example of a TV-SDP

TV-SDP cannot be greater than 0.93. Moreover, we can get arbitrarily close to the exact optimal value of the TV-SDP by increasing the degree of the candidate polynomial solutions (cf. Section 2.3.1) or the level in the hierarchy of our dual problems (cf. Section 2.4.1).

2.1.1 Related literature

Time-varying SDPs contain as special case the time-varying versions of most common classes of convex optimization problems, including linear programs, convex quadratic programs, and second-order cone programs. In the linear programming case, this problem has been studied in the literature under the name of continuous linear programs (CLPs). In its most general form, a CLP is a problem of the type

$$\begin{array}{ll} \sup\limits_{x(t)} & \int_0^1 \langle c(t), x(t)\rangle dt \\ \text{subject to} & A(t)x(t) + \int_0^t D(t,s)x(s)ds \leq b(t) \quad \forall t \in [0,1] \text{ a.e.,} \end{array} \tag{2.4}$$

where $A(t), D(t,s) \in \mathbb{R}^{m\times n}, c(t) \in \mathbb{R}^n$, and $b(t) \in \mathbb{R}^m$, for all $t, s \in [0,1]$.

This problem was introduced by Bellman [36] and has since been studied by several authors who have provided algorithms, structural results, or a duality theory for CLPs; see e.g. [127, 200, 16, 130, 201, 86, 50, 162, 21, 17, 135, 188, 78, 207] and references therein. Several applications, e.g. in manufacturing, transportation, robust optimization, queueing theory, and revenue management, can also be found in these references.

Since CLPs are perceived as a hard problem class in general, most authors make additional assumptions on how the problem data varies with time, or, in the case of the so-called "separated CLPs" (SCLPs), how the kernel terms and the non-kernel terms interact [170, 168, 171, 135, 78, 207, 50, 162, 17]. SLCPs enjoy many properties that general CLPs do not. For instance, under mild assumptions, SCLPs with piecewise-polynomial data admit piecewise-polynomial solutions [169]—an attractive feature from an algorithmic point view. Unfortunately, the situation for TV-SDPs is not as nice, even without a kernel term. For example, consider problem (2.1) with $n = 2$

and data

$$A_0(t) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, A_1(t) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, A_2(t) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, D_1(t,s) = D_2(t,s) = 0, \text{ and } c(t) = \begin{pmatrix} t \\ 1-t \end{pmatrix}.$$

This TV-SDP has no kernel terms. Furthermore, all its data is constant except for the objective function which varies linearly with time. Its unique optimal solution, however, is easily seen to be $\frac{c(t)}{\|c(t)\|}$, i.e.,

$$x(t) = \frac{1}{\sqrt{t^2 + (1-t)^2}} \begin{pmatrix} t \\ 1-t \end{pmatrix},$$

which is not a piecewise-polynomial function.

The closest work in the CLP literature to our work is the paper [33] by Bampou and Kuhn. The authors of this paper also assume that the data of the their CLP varies polynomially with time and employ semidefinite programming to approximate the optimal solution by polynomial (and piecewise polynomial) functions of time. Our approach here generalizes their nice algorithms and convergence guarantees to the SDP setting. In [33], the authors also make use of the rich duality theory of CLPs to get a sequence of upper bounds that converges to the optimal value of (2.4) under certain conditions. The duality framework that we present in this chapter is different in nature and is closer in spirit to the approach in [123], [32]. As it turns out, it suffices for us to assume boundedness of the primal feasible set to guarantee convergence of our dual bounds to the optimal value of the TV-SDP.

The only generalization of continuous linear programs that we are aware of appears in the work of Wang, Zhang, and Yao in [205], which makes a number of important contributions to separated continuous conic programs. The assumptions in [205] are however stronger than the ones we make here. In particular, there are separation assumptions on the kernel and non-kernel terms in [205] and the data to the problem is assumed to vary only linearly with time. Another work related to this chapter is the work by Lasserre in [123], which studies a parametric polynomial optimization problem of the form

$$\begin{array}{ll} \sup\limits_{x(y)\in\mathbb{R}^n} & \int_{y\in K} f(x(y),y)\mathrm{d}\phi(y) \\ \text{subject to} & h_j(x(y),y) \geq 0 \qquad \forall j \in \{1,\ldots,r\}, \forall y \in K \ \phi\text{-a.e.}, \end{array} \tag{2.5}$$

where ϕ is a probability distribution on some compact basic semialgebraic set $K \subseteq \mathbb{R}^s$, and $h_j(x,y)$ are polynomial functions of x and y. An inequality involving y is valid ϕ-a.e. if it is valid for all y in K except on some set K' with $\phi(K') = 0$. When the kernel terms in (2.2) are zero, problem (2.1) can in theory be put in the form of (2.5) by setting $s = 1$ and replacing the semidefinite constraint with nonnegativity of all $2^m - 1$ polynomials that form the principal minors of $Fx(t)$. Our duality framework in Section 2.4 is inspired by the approach in [123]. However, as we are dealing with a much more structured problem, we are also able to find the best polynomial solution of a given degree to (2.1) with an SDP of tractable size, as well as prove asymptotic optimality of polynomial solutions even in presence of the kernel terms.

Finally, we remark that at a broader level, the idea of using semidefinite programming to find polynomial solutions (or "policies") to dynamic or uncertain optimization problems has been applied before to questions in multi-stage robust and stochastic optimization; see e.g. [40] and [32].

2.1.2 Organization and contributions of the chapter

This chapter is organized as follows. In Section 2.2, we prove that under a boundedness assumption, the optimal value of the TV-SDP in (2.1) is attained (Theorem 3). This proof is obtained by combining two theorems that are used also in other sections of the chapter. The first (Theorem 1) shows that a sequence of linear functionals that satisfies a certain boundedness property on nonnegative polynomials has a weakly convergent subsequence. The second (Theorem 2) shows that when a weakly convergent sequence of functions in $\mathbf{L^n}$ satisfies linear inequalities of the type in (2.3), then so does its weak limit.

In Section 2.3, we prove that under a strict feasibility assumption, polynomial solutions are arbitrarily close to being optimal to the TV-SDP in (2.1) (Theorem 4). We also show that this assumption cannot be removed in general (Example 1). Furthermore, we show how sum of squares techniques combined with certain matrix Positivstellensatzë enable the search for the best polynomial solution of a given degree to be cast as an SDP of polynomial size (Theorem 6).

In Section 2.4, we develop a hierarchy of dual problems (or relaxations) that give a sequence of improving upper bounds on the optimal value of the TV-SDP in (2.1). We show that under a boundedness assumption, these upper bounds converge to the optimal value of the TV-SDP (Theorem 7). We also show that our dual problems can be cast as SDPs (Theorem 8). For a given TV-SDP, the dimensions of the matrices that feature in both our primal and dual SDP hierarchies grow only linearly with the order of the hierarchy.

In Section 2.5, we present applications of time-varying semidefinite programs to a maximum-flow problem with time-varying edge capacities, a wireless coverage problem with time-varying coverage requirements, and to bi-objective semidefinite optimization where the goal is to approximate the Pareto curve in one shot. Finally, we end with some future research directions in Section 2.6.

2.1.3 Notation

We denote

- the $(i,j)^{th}$ entry of a matrix A by A_{ij},
- the trace of a matrix A by $\mathrm{Tr}(A)$,
- the vector of all ones by $\mathbf{1}$,
- the identity matrix by I,
- the diagonal matrix with the vector $x \in \mathbb{R}^n$ on its diagonal by $diag(x)$,

- the standard inner product in $\mathbb{R}^n$ by $\langle \cdot, \cdot \rangle$; i.e., for two vectors $x, y \in \mathbb{R}^n$, $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$,
- the infinity-norm of a vector by $\| \cdot \|_\infty$; i.e., for a vector $x \in \mathbb{R}^n$, $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$,
- the set of $n \times n$ (constant) symmetric matrices by $\mathcal{S}^n$ and its subset of positive semidefinite matrices by $\mathcal{S}^{n+}$,
- the degree of a polynomial p by $\deg(p)$ (when p is a vector of polynomials, $deg(p)$ denotes the maximum degree of its entries),
- the set of $n \times m$ matrices whose components are polynomials in the variable t with real coefficients by $\mathbb{R}^{n\times m}[t]$. For $d \in \mathbb{N}$, $\mathbb{R}_d^{n\times m}[t]$ denotes the subset of $\mathbb{R}^{n\times m}[t]$ consisting of matrices whose entries are polynomials of degree at most d. When $m = 1$, we simply use the notation $\mathbb{R}^n[t]$ and $\mathbb{R}_d^n[t]$, and when $n = 1$ as well, we simplify the notation to $\mathbb{R}[t]$ and $\mathbb{R}_d[t]$.
- We denote the set of linear functionals on $\mathbb{R}^n[t]$ by $\mathbf{M}^n$.
- For $\mu \in \mathbf{M}^n$, we denote by $\mu_i : \mathbb{R}[t] \to \mathbb{R}$ the unique linear functional that satisfies
$$\mu(g) = \sum_{i=1}^n \mu_i(g_i) \quad \forall g \in \mathbb{R}^n[t].$$
- For a function $f \in \mathbf{L^n}$, we denote by l_f the element of $\mathbf{M}^n$ defined by
$$l_f(g) := \langle f, g \rangle_{\mathbf{L^n}} = \int_0^1 \langle f(t), g(t) \rangle \mathrm{d}t \quad \forall g \in \mathbb{R}^n[t].$$

2.2 The Optimal Value of a Bounded TV-SDP is Attained

In this section, we study the following question: If the optimal value *opt* of (2.1) is finite (i.e., the problem is feasible and bounded above), does there exist a function $x^* \in \mathbf{L^n}$ such that $\langle c, x^* \rangle_{\mathbf{L^n}} = opt$? Many of the arguments given here will be used again in Section 2.4 on duality theory.

The question of attainment of the optimal value (i.e., existence of solutions) is a very basic one and has been studied in the continuous linear programming literature already; see e.g. [86]. In the TV-SDP case, note that even for standard SDPs that do not depend on time, the optimal value is not always attained unless the feasible set is bounded. We prove in this section that under the following boundedness assumption

$$\text{“}\exists \gamma > 0 \text{ such that for all feasible solutions } x \text{ to (2.1)}, \|x(t)\|_\infty \leq \gamma \text{ for all } t \in [0,1] \text{ a.e.”}, \tag{2.6}$$

the optimal value of the TV-SDP in (2.1) is always attained. This is not an immediate fact as the search space $\mathbf{L^n}$ is infinite dimensional. The idea is to prove that a sequence

of feasible solutions to a TV-SDP whose objective value approaches the optimal value must have a converging subsequence and that the limit of the subsequence must also be feasible. It turns out that the right notion of convergence in this context is *weak convergence*. We begin by stating the definition, and then prove that the weak limit of a sequence of feasible solutions is again feasible.

Definition 2.2.1 (Weak Convergence)**.** *A sequence of linear functionals* $\{\mu^i\}$ *in* $\mathbf{M}^n$ *converges weakly to a linear functional* $\mu^\infty \in \mathbf{M}^n$ *(we write* $\mu^i \rightharpoonup \mu^\infty$*) if for all* $p \in \mathbb{R}^n[t]$,

$$\mu^i(p) \to \mu^\infty(p) \text{ as } i \to \infty.$$

Similarly, a sequence of functions $\{f^i\}$ *in* $\mathbf{L^n}$ *converges weakly to a function* $f^\infty \in \mathbf{L^n}$ *(we write* $f^i \rightharpoonup f^\infty$*) if* $l_{f^i} \rightharpoonup l_{f^\infty}$ *as* $i \to \infty$.

The next theorem shows a compactness result for the set $\mathbf{M}^n$.

Theorem 1. *Let* $\{\mu^d\}$ *be a sequence of linear functionals in* $\mathbf{M}^n$*. If the following implication holds for every* $d \in \mathbb{N}$ *and every polynomial* $q \in \mathbb{R}^n_d[t]$*:*

$$q_i(t) \geq 0 \quad \forall t \in [0,1],\ \forall i \in \{1\ldots,n\} \implies |\mu^d(q)| \leq \sum_{i=1}^{n} \int_0^1 q_i(t)\mathrm{d}t,$$

then there exists a function $f \in \mathbf{L^n}$ *and a subsequence of* $\{\mu^d\}$ *that converges weakly to* l_f*.*

In the proof of this theorem, we will invoke the following lemma, which is obtained by a direct application of a result of Lasserre [125, Theorem 3.12a].[1]

Lemma 2.2.2 (See Theorem 3.12a in [125])**.** *For a linear functional* $\lambda \in \mathbf{M}^1$*, if there exists a scalar* κ *such that the inequalities*

$$0 \leq \lambda(h) \leq \kappa \int_0^1 h(t)\ \mathrm{d}t$$

hold for every polynomial $h \in \mathbb{R}[t]$ *that is nonnegative on* $[0,1]$*, then there exists a function* $f \in \mathbf{L^1}$ *such that* $\lambda(g) = l_f(g), \forall g \in \mathbb{R}[t]$.

Theorem 1. The ideas of the proof are inspired by those in [204, Chap. 7]. Let $\{b_0, b_1, \ldots\}$ be a basis of $\mathbb{R}^n[t]$ where all entries of the polynomials b_j are of the form t^k for some nonnegative integer k. Let d_j denote the maximum degree of the entries of b_j. It is clear by assumption that $|\mu^i(b_j)| \leq n$ for every $i, j \in \mathbb{N}$ such that $i \geq d_j$. Consider the sequence of real numbers $\{\mu^i(b_0)\}_{i \geq d_0}$. This sequence is bounded in absolute value by n. As such, it has a convergent subsequence $\{\mu^{i,(0)}(b_0)\}$. Next consider $\{\mu^{i,(0)}(b_1)\}_{i \geq d_1}$. Again, this is a sequence of real numbers that is bounded in absolute value by n and so it has a convergent subsequence $\{\mu^{i,(1)}(b_1)\}$. Iterating

[1] To get the statement of the lemma, apply [125, Theorem 3.12a] with $n = 1, \mathbb{K} = [0,1], m = 2, g_1 = t, g_2 = 1-t, L_y = \lambda, L_z = l_v$, where $v \in \mathbf{L^1}$ is the constant function equal to one, and observe that for any $p \in \mathbb{R}[t]$, the polynomials $p^2 g_1, p^2 g_2, p^2 g_1 g_2$ are nonnegative on the interval $[0,1]$.

this procedure, we obtain, for each integer $r \geq 0$, a subsequence of linear functionals $\{\mu^{i,(r)}\}$ with the property that $\{\mu^{i,(r+1)}\} \subseteq \{\mu^{i,(r)}\}$. Moreover, for all $j, r \in \mathbb{N}$ with $r \geq d_j$, the sequence of numbers $\{\mu^{i,(r)}(b_j)\}$ converges as $i \to \infty$. Now consider the diagonal sequence of linear functionals $\{\mu^{i,(i)}\}$. For every j, $\{\mu^{i,(i)}(b_j)\}$ converges as $i \to \infty$ as the sequence of linear functionals $\{\mu^{i,(i)}\}_{i \geq d_j}$ is a subsequence of $\{\mu^{i,(d_j)}\}$. Since the functions $\{b_i\}$ span $\mathbb{R}^n[t]$ and the elements of the sequence $\{\mu^{i,(i)}\}$ are linear functionals, the sequence $\{\mu^{i,(i)}(g)\}$ converges for all polynomial functions $g \in \mathbb{R}^n[t]$. Let μ^∞ be the linear functional defined by

$$\mu^\infty(g) = \lim_{i \to \infty} \mu^{i,(i)}(g) \quad \forall g \in \mathbb{R}^n[t]. \tag{2.7}$$

We have just proven that the sequence $\{\mu^{i,(i)}\}$ converges weakly to μ^∞. The claim of the theorem would be established if we show that there exists a function $f \in \mathbf{L^n}$ such that $\mu^\infty(g) = l_f(g), \forall g \in \mathbb{R}^n[t]$. In order to get this statement from Lemma 2.2.2, for $j \in \{1, \ldots, n\}$, let $\lambda_j \in \mathbf{M^1}$ be defined as

$$\lambda_j(w) := \int_0^1 w(t)\, \mathrm{d}t - \mu_j^\infty(w) \quad \forall w \in \mathbb{R}[t].$$

Let $h \in \mathbb{R}[t]$ be a polynomial that is nonnegative on $[0, 1]$. Take $h^{(j)} \in \mathbb{R}^n[t]$ to be the vector-valued polynomial whose entries are all identically zero except for the j^{th} one that is equal to h. From (2.7) we see that

$$\mu_j^\infty(h) = \mu^\infty(h^{(j)}) = \lim_{i \to \infty} \mu^{i,(i)}(h^{(j)}) = \lim_{i \to \infty} \mu_j^{i,(i)}(h).$$

Since for i larger than the degree of h,

$$|\mu_j^{i,(i)}(h)| = |\mu^{i,(i)}(h^{(j)})| \leq \sum_{k=1}^{n} \int_0^1 h_k^{(j)}(t)\, \mathrm{d}t = \int_0^1 h(t)\, \mathrm{d}t,$$

we have that that $|\mu_j^\infty(h)| \leq \int_0^1 h(t)\, \mathrm{d}t$, and therefore

$$|\lambda_j(h)| \leq \left| \int_0^1 h(t)\, \mathrm{d}t \right| + |\mu_j^\infty(h)| \leq 2 \int_0^1 h(t)\, \mathrm{d}t.$$

Similarly, it is straightforward to argue that $\lambda_j(h) \geq 0$. Hence, by Lemma 2.2.2, for each $j \in \{1, \ldots, n\}$, there exists a function $\hat{f}_j \in \mathbf{L^1}$ such that $\lambda_j(w) = l_{\hat{f}_j}(w), \forall w \in \mathbb{R}[t]$. Therefore,

$$\mu_j^\infty(w) = \int_0^1 \big(1 - \hat{f}_j(t)\big) w(t)\, \mathrm{d}t, \forall w \in \mathbb{R}[t].$$

The function $f \in \mathbf{L^n}$ that we were after can hence be taken to be $f := (1 - \hat{f}_1, \ldots, 1 - \hat{f}_n)^T$. □

The next theorem shows that when all functions in a sequence satisfy linear inequalities of the type in (2.3), their weak limit does the same.

Theorem 2. *Let the operator F be as in (2.2). If a sequence of functions $\{f_k\}$ in $\mathbf{L^n}$ converges weakly to a function $f_\infty \in \mathbf{L^n}$ and satisfies $Ff_k \in \mathbf{S^{m+}}([0,1])$ for all $k \in \mathbb{N}$, then $Ff_\infty \in \mathbf{S^{m+}}([0,1])$.*

To prove this theorem, we need the following lemma, which also implies that the set $\mathbf{S^{m+}}([0,1])$ is self-dual. This is a generalization of the corresponding statement for the non time-varying case, which states that the cone $\mathcal{S}^{n+}$ is self-dual.

Lemma 2.2.3. *For any function $Q \in \mathbf{S^m}$, $Q \in \mathbf{S^{m+}}([0,1])$ if and only if*

$$\langle Q, P\rangle_{\mathbf{S^m}} \geq 0 \text{ for all } P \in \mathbf{S^{m+}}([0,1]) \cap \mathbb{R}^{m\times m}[t].$$

Proof. The only if part is straightforward. For the other direction, fix $Q \in \mathbf{S^m}$ and assume that $\langle Q, P\rangle_{\mathbf{S^m}} \geq 0$ for all $P \in \mathbf{S^{m+}}([0,1]) \cap \mathbb{R}^{m\times m}[t]$. For $t \in [0,1]$, let $\lambda(t)$ be the smallest eigenvalue of $Q(t)$ and $u(t)$ be an associated eigenvector of norm one. Denote by $1_{\lambda(t)<0}$ the univariate function over $t \in [0,1]$ that is equal to 1 when $\lambda(t) < 0$ and zero otherwise. Let $P^\infty(t) := 1_{\lambda(t)<0}u(t)u(t)^T$. We claim that $\langle Q, P^\infty\rangle_{\mathbf{S^m}} \geq 0$. This would imply that

$$\int_0^1 1_{\lambda(t)<0}\lambda(t)dt = \langle Q, P^\infty\rangle_{\mathbf{S^m}} \geq 0,$$

which proves that $\lambda(t)$ is nonegative almost everywhere on $[0,1]$; i.e., the desired result.

To prove the claim, observe that since continuous functions are dense in the space of bounded and measurable functions on $[0,1]$ (see e.g. [1, Theorem 2.19]), for every positive integer k, there exist continuous functions $\phi_k : [0,1] \to \mathbb{R}$ and $u_k : [0,1] \to \mathbb{R}^n$ such that

$$\int_0^1 (\phi_k(t) - 1_{\lambda(t)<0})^2 \mathrm{d}t \leq \frac{1}{k} \text{ and } \int_0^1 \|u_k(t) - u(t)\|_\infty^2 \mathrm{d}t \leq \frac{1}{k}.$$

Notice that without loss of generality we can assume that for all $k \in \mathbb{N}$ and $t \in [0,1]$ we have $\phi_k(t) \geq 0$ as

$$\big||\phi_k(t)| - 1_{\lambda(t)<0}\big| \leq |\phi_k(t) - 1_{\lambda(t)<0}|.$$

The Stone-Weierstrass theorem (see e.g. [198]) can now be utilized to conclude that for every positive integer k, there exist polynomial functions $\tilde{\phi}_k : [0,1] \to \mathbb{R}$, $\tilde{u}_k : [0,1] \to \mathbb{R}^n$ such that

$$0 \leq \tilde{\phi}_k(t) - \phi_k(t) \leq \frac{1}{k} \text{ and } \|\tilde{u}_k(t) - u_k(t)\|_\infty^2 \leq \frac{1}{k} \quad \forall t \in [0,1].$$

We can thus assume without loss of generality again that the functions ϕ_k and u_k are polynomial functions of the variable t.

Now let $P^k(t) = \phi_k(t)u_k(t)u_k(t)^T$. Then (i) $P^k \in \mathbf{S^m}^+([0,1]) \cap \mathbb{R}^{m\times m}[t]$, and (ii) $\|P^\infty - P^k\|_{\mathbf{S^m}} \to 0$ as $k \to \infty$, where $\|.\|_{\mathbf{S^m}}$ here denotes the norm associated to the scalar product $\langle ., . \rangle_{\mathbf{S^m}}$. From the Cauchy-Schwarz inequality we have

$$|\langle Q, P^\infty\rangle_{\mathbf{S^m}} - \langle Q, P^k\rangle_{\mathbf{S^m}}| \leq \|Q\|_{\mathbf{S^m}}\|P^\infty - P^k\|_{\mathbf{S^m}}.$$

As (i) implies that $\langle Q, P^k\rangle_{\mathbf{S^m}} \geq 0$ for all k, and (ii) implies that the right hand side of the above inequality goes to zero as k goes to infinity, we conclude that $\langle Q, P^\infty\rangle_{\mathbf{S^m}} \geq 0$. □

Theorem 2. For an element $y \in \mathbf{L^n}$, we denote by $\tilde{y}$ the element of $\mathbf{L^{n+1}}$ defined by $\tilde{y} := \begin{pmatrix} 1 \\ y \end{pmatrix}$. By applying Fubini's double integration theorem on the region $\{(t,s) \in [0,1]^2 \mid s \leq t\}$, it is straightforward to see that

$$\langle Fy, P\rangle_{\mathbf{S^m}} = \langle \tilde{y}, F^*P\rangle_{\mathbf{L^{n+1}}} \quad \forall y \in \mathbf{L^n},\ \forall P \in \mathbf{S^m},$$

where F^* is the adjoint of the affine operator F (see equation (2.14) in Section 2.4 for its explicit expression). Now fix a function $P \in \mathbf{S^m}^+([0,1]) \cap \mathbb{R}^{m\times m}[t]$. Using the easy direction of Lemma 2.2.3 and the fact that $Ff_k \in \mathbf{S^m}^+([0,1])$ for all k, we have that $\langle Ff_k, P\rangle_{\mathbf{S^m}} \geq 0$ for all k. This implies that $\langle \tilde{f}_k, F^*P\rangle_{\mathbf{L^{n+1}}} \geq 0$ for all k. By weak convergence, we conclude that $\langle \tilde{f}_\infty, F^*P\rangle_{\mathbf{L^{n+1}}} \geq 0$, implying in turn that $\langle Ff_\infty, P\rangle_{\mathbf{S^m}} \geq 0$. Since P was arbitrary in $\mathbf{S^m}^+([0,1]) \cap \mathbb{R}^{m\times m}[t]$, using Lemma 2.2.3 again, we have $Ff_\infty \in \mathbf{S^m}^+([0,1])$. □

We are now ready to show that a bounded TV-SDP attains its optimal value.

Theorem 3. *If the TV-SDP in (2.1) is feasible and satisfies the boundedness assumption in (2.6), then there exists a feasible function $x^{opt} \in \mathbf{L^n}$ that attains its optimal value.*

Proof. Let opt denote the optimal value of (2.1), which is finite under the assumptions of the theorem. From (2.6), there exists a scalar $\gamma > 0$ such that any feasible solution $x \in \mathbf{L^n}$ to the TV-SDP satisfies $\|x(t)\|_\infty \leq \gamma$ for all $t \in [0,1]$ a.e.. Hence, for any positive integer k, there exists a feasible solution $x^k \in \mathbf{L^n}$, with $\|x^k(t)\|_\infty \leq \gamma\ \forall t \in [0,1]$ a.e., such that

$$\langle c, x^k\rangle_{\mathbf{L^n}} \geq \text{opt} - \frac{1}{k}. \tag{2.8}$$

Let us now consider the sequence of linear functionals $\left\{\frac{l_{x^k}}{\gamma}\right\}$, which satisfies the conditions of Theorem 1. Therefore, a subsequence of the functions $\{x^k\}$ converges weakly to a limit $x^\infty \in \mathbf{L^n}$. It is clear by weak convergence that x^∞ achieves the optimal value to (2.1), and Theorem 2 guarantees that x^∞ is feasible to the TV-SDP. Letting $x^{\text{opt}} = x^\infty$ gives the desired result. □

2.3 The Primal Approach: Polynomial Solutions to a TV-SDP

We switch our focus in this section to algorithmic questions. We show in Section 2.3.2 that when the data $c, A_0, \ldots, A_n, D_1, \ldots, D_n$ to our TV-SDP belongs to the class of polynomial functions, then the best polynomial solution of a given degree to the TV-SDP can be found by solving a semidefinite program of tractable size. This motivates us to study whether one can always find feasible solutions to a TV-SDP that are arbitrarily close to being optimal just by searching over polynomial functions. While this is not always true (see Example 1 below), in Section 2.3.1 we show that it is true under a strict feasibility assumption (see Definition 2.3.1).

Example 1. *Consider the TV-SDP in (2.1) with $n = 1$,*

$$c(t) = 0, A_0(t) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} - t & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, A_1(t) = \begin{pmatrix} t - \frac{1}{2} & 0 & 0 & 0 \\ 0 & t - \frac{1}{2} & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \text{ and } D_1 = 0.$$

The resulting constraints read

$$\left(t - \frac{1}{2}\right) x(t) \geq 0, \ \left(t - \frac{1}{2}\right) (x(t) - 1) \geq 0, \ 0 \leq x(t) \leq 1 \quad \forall t \in [0, 1] \text{ a.e.}.$$

The unique feasible solution $x^{opt}(t)$ to this TV-SDP, up to a set of measure zero, is

$$x^{opt}(t) = \begin{cases} 0, & \text{if } t \leq \frac{1}{2}, \\ 1, & \text{if } t > \frac{1}{2}. \end{cases}$$

It is clear that x^{opt} is not continuous, let alone polynomial.

For the remainder of this chapter, for a set $S \in \{\mathbf{S^m}, \mathbf{S^{m+}}([0,1])\}$ and a non-negative integer d, we define S_d to be the set of functions $x \in S$ whose entries are polynomials of degree d, i.e.

$$\mathbf{S^m}_d = \mathbf{S^m} \cap \mathbb{R}^{m \times m}_d[t], \quad \mathbf{S^{m+}}([0,1])_d = \mathbf{S^{m+}}([0,1]) \cap \mathbb{R}^{m \times m}_d[t].$$

2.3.1 Polynomials are optimal under a strict feasibility assumption

We show in this section that under the following strict feasibility assumption, the optimal value of the TV-SDP in (2.1) remains the same when the function class $\mathbf{L^n}$ is replaced with $\mathbb{R}^n[t]$.

Definition 2.3.1. *We say that the TV-SDP in (2.1) is strictly feasible if there exists a function $x^s \in \mathbf{L^n}$ and a positive scalar ε such that*

$$Fx^s(t) \succeq \varepsilon I \quad \forall t \in [0, 1] \text{ a.e.}.$$

Theorem 4. *Consider the TV-SDP in (2.1) with its optimal value denoted by opt. If the TV-SDP is strictly feasible, then there exists a sequence of feasible polynomial solutions $\{x^k\}$ such that*

$$\langle c, x^k \rangle_{\mathbf{L^n}} \to opt \text{ as } k \to \infty.$$

As we will shortly see in the proof, the strict feasibility assumption enables us to approximate any feasible solution of (2.1) by a continuous, and later polynomial, solution. We use *mollifying operators* to obtain the continuous approximation.

Definition 2.3.2 (See [1])**.** *The* mollifying operator $\mathcal{M}_v : \mathbf{L^1} \to \mathbf{L^1}$*, indexed by a nonnegative integer v, is the linear operator defined by*

$$(\mathcal{M}_v f)(t) = \int_0^1 vJ(v(t-s))f(s)\mathrm{d}s \quad \forall f \in \mathbf{L^1}$$

where $J(t) = c\exp(-\frac{1}{1-t^2})$ when $t \in [-1,1]$ and $J(t) = 0$ otherwise, and c is so that $\int_{\mathbb{R}} J(t)\ \mathrm{d}t = 1$.

Remark 1. *To lighten our notation, we write $\mathcal{M}_v f(t)$ instead of $(\mathcal{M}_v f)(t)$. We also remark that one can extend the definition of mollifying operators to functions that are not scalar valued by making them act element-wise. For example, the extension to spaces $\mathbf{L^n}$ and $\mathbf{S^m}$ would be defined as follows:*

$$\mathcal{M}_v f := (\mathcal{M}_v f_i)_i \ \forall f \in \mathbf{L^n} \text{ and } \mathcal{M}_v P := (\mathcal{M}_v P_{ij})_{ij} \ \forall P \in \mathbf{S^m}.$$

Any property of mollifying operators that we prove on scalar-valued functions below extends in a straightforward manner to functions that are vector or matrix valued.

Proposition 1 (See Theorem 2.29 in [1])**.** *For all $f \in \mathbf{L^1}$ and all $v \in \mathbb{N}$, the function $\mathcal{M}_v f$ is continuous. Moreover,*

$$\int_0^1 |\mathcal{M}_v f(t) - f(t)|\mathrm{d}t \to 0 \text{ as } v \to \infty.$$

Furthermore, if f is a continuous function of t, then

$$\sup_{t\in[0,1]} |\mathcal{M}_v f(t) - f(t)| \to 0 \text{ as } v \to \infty.$$

Lemma 2.3.3. *For for any $v \in \mathbb{N}$, the mollifying operator $\mathcal{M}_v$ satisfies the following properties:*

(a) For any $M \in \mathbf{S^m}$, if $M(t) \succeq 0\ \forall t \in [0,1]$ a.e., then $\mathcal{M}_v M(t) \succeq 0\ \forall t \in [0,1]$.

(b) For any $a \in \mathbb{R}[t]$ and $x \in \mathbf{L^1}$, $\sup_{t\in[0,1]} |a(t)\mathcal{M}_v x(t) - \mathcal{M}_v(a \cdot x)(t)| \to 0$ as $v \to \infty$.

(c) For any polynomial function $d : \mathbb{R}^2 \to \mathbb{R}$ and $x \in \mathbf{L^1}$, let $g(t) = \int_0^t d(t,s)x(s)\mathrm{d}s$. Then

$$\sup_{t\in[0,1]} \left| \mathcal{M}_v g(t) - \int_0^t d(t,s)\mathcal{M}_v x(s)\mathrm{d}s \right| \to 0 \text{ as } v \to \infty.$$

Proof. The proof of (a) simply follows from the fact that the function J is nonnegative on $\mathbb{R}$.

To prove (b), let $L := \sup_{t\in[0,1]} |a'(t)|$ and $\gamma := \sup_{t\in[0,1]} |x(t)|$. Notice that

$$a(t)\mathcal{M}_v x(t) - \mathcal{M}_v(a\cdot x)(t) = \int_0^1 vJ(v(t-s))(a(t)-a(s))x(s)\mathrm{d}s.$$

Hence,

$$|a(t)\mathcal{M}_v x(t) - \mathcal{M}_v(ax)(t)| \leq L\gamma \int_0^1 vJ(v(t-s))|t-s|\mathrm{d}s.$$

By the change of variable $u := v(s-t)$ and in view of the evenness of the function J, we get

$$\int_0^1 vJ(v(t-s))|t-s|\mathrm{d}s = \frac{1}{v}\int_{-vt}^{v(1-t)} J(u)\,|u|\,\mathrm{d}u \leq \frac{1}{v}\int_{-1}^{1} J(u)\mathrm{d}u \leq \frac{1}{v}.$$

Therefore,

$$\sup_{t\in[0,1]} |a(t)\mathcal{M}_v x(t) - \mathcal{M}_v(a\cdot x)(t)| \leq \frac{L\gamma}{v}$$

and the claim follows.

Let us now prove (c). Observe that, on the one hand, for every $t\in[0,1]$,

$$\begin{aligned}\left|\mathcal{M}_v g(t) - \int_0^t d(t,s)\mathcal{M}_v x(s)\,\mathrm{d}s\right| &\leq |\mathcal{M}_v g(t) - g(t)| + \left|g(t) - \int_0^t d(t,s)\mathcal{M}_v x(s)\mathrm{d}s\right| \\ &\leq |\mathcal{M}_v g(t) - g(t)| + \int_0^t |d(t,s)|\cdot|x(s) - \mathcal{M}_v x(s)|\,\mathrm{d}s \\ &\leq |\mathcal{M}_v g(t) - g(t)| + \sup_{t,s\in[0,1]} |d(t,s)| \int_0^1 |x(s) - \mathcal{M}_v x(s)|\,\mathrm{d}s.\end{aligned}$$

On the other hand, from Proposition 1 (and continuity of g), we know that

$$\sup_{t\in[0,1]} |\mathcal{M}_v g(t) - g(t)| \to 0 \text{ as } v\to\infty,$$

and

$$\int_0^1 |\mathcal{M}_v x(s) - x(s)|\,\mathrm{d}s \to 0 \text{ as } v\to\infty.$$

Combining these three facts, we conclude that

$$\sup_{t\in[0,1]} |\mathcal{M}_v g(t) - \int_0^t d(t,s)\mathcal{M}_v x(s)\mathrm{d}s| \to 0 \text{ as } v\to\infty.$$

□

Properties (b) and (c) in Lemma 2.3.3 give the following corollary.

Corollary 1. *Let F be as in (2.2) with $A_0, \ldots, A_n, D_1, \ldots D_n$ polynomial and let $\|\cdot\|$ be any matrix norm. Then,*

$$\sup_{t\in[0,1]} \|\mathcal{M}_v F x(t) - F\mathcal{M}_v x(t)\| \to 0 \text{ as } v \to \infty.$$

We now go back to the proof of optimality of polynomial solutions. The idea is as follows. We know that for any $\varepsilon > 0$ there exists a feasible solution x^ε whose objective value is within ε of opt when opt $< \infty$ and larger than $\frac{1}{\varepsilon}$ when opt $= +\infty$. We construct a sequence of feasible polynomial solutions whose objective value converge to $\langle c, x^\varepsilon\rangle_{\mathbf{L^n}}$. We do so in three steps. First, using existence of a strictly feasible solution x^s to (2.1), we perturb x^ε slightly to make it strictly feasible without changing its objective value by much. Second, we approximate the perturbed solution by a continuous solution using mollifying operators. Finally, we invoke the Stone-Weierstrass theorem to approximate the continuous solution with a polynomial solution.

of Theorem 4. For any $\varepsilon > 0$, let x^ε be a feasible solution to the TV-SDP in (2.1) such that

$$\text{opt} - \langle c, x^\varepsilon\rangle_{\mathbf{L^n}} \leq \varepsilon \text{ if opt} < \infty \text{ and } \langle c, x^\varepsilon\rangle_{\mathbf{L^n}} \geq \frac{1}{\varepsilon} \text{ if opt} = +\infty.$$

Let x^s be any strictly feasible solution to the TV-SDP in (2.1) and for $\lambda \in (0,1)$ let

$$x^{\lambda,\varepsilon} := (1-\lambda)x^\varepsilon + \lambda x^s.$$

Observe that for all $\varepsilon > 0$ and $\lambda \in (0,1)$ the function $x^{\lambda,\varepsilon}$ is also strictly feasible to (2.1). Moreover, as $\lambda \to 0$, $\langle c, x^{\lambda,\varepsilon}\rangle_{\mathbf{L^n}} \to \langle c, x^\varepsilon\rangle_{\mathbf{L^n}}$.

For a nonnegative integer v, let $\mathcal{M}_v$ be the mollifying operator that appears in Definition 2.3.2 and Remark 1. For all $\varepsilon > 0$, $\lambda \in (0,1)$, and $v \in \mathbb{N}$, let $x^{v,\lambda,\varepsilon} := \mathcal{M}_v x^{\lambda,\varepsilon}$. The function $x^{v,\lambda,\varepsilon}$ is continuous by Proposition 1. We claim that $x^{v,\lambda,\varepsilon}$ is strictly feasible to the TV-SDP in (2.1) for any $\varepsilon > 0$ and $\lambda \in (0,1)$ when v is large enough. Indeed, for any such ε and λ, there exists $\beta_{\lambda,\varepsilon} > 0$ such that $Fx^{\lambda,\varepsilon}(t) \succeq \beta_{\lambda,\varepsilon} I \ \forall t \in [0,1]$ a.e.. By property (a) of Lemma 2.3.3,

$$\mathcal{M}_v F x^{\lambda,\varepsilon}(t) \succeq \beta_{\lambda,\varepsilon} I \ \forall t \in [0,1].$$

Let $\|\cdot\|$ be any matrix norm. Using Corollary 1,

$$\sup_{t\in[0,1]} \|\mathcal{M}_v F x^{\lambda,\varepsilon}(t) - F x^{v,\lambda,\varepsilon}(t)\| \to 0 \text{ as } v \to \infty.$$

By continuity of the minimum eigenvalue function we conclude that for v high enough,

$$F x^{v,\lambda,\varepsilon}(t) \succeq \frac{\beta_{\lambda,\varepsilon}}{2} I \ \forall t \in [0,1].$$

Moreover, for all $\varepsilon > 0$, $\lambda \in (0,1)$, Proposition 1 implies that

$$\langle c, x^{v,\lambda,\varepsilon}\rangle_{\mathbf{L^n}} \to \langle c, x^{\lambda,\varepsilon}\rangle_{\mathbf{L^n}} \text{ as } v \to \infty.$$

As a final step, for a fixed $\varepsilon > 0$, $\lambda \in (0,1)$, and $v \in \mathbb{N}$, we invoke the Stone-Weierstrass theorem to approximate $x^{v,\lambda,\varepsilon}$ by a sequence $\{p^{s,v,\lambda,\varepsilon}\}_{s\in\mathbb{N}}$ of polynomial elements of $\mathbf{L^n}$ such that

$$\sup_{t\in[0,1]} \|x^{v,\lambda,\varepsilon}(t) - p^{s,v,\lambda,\varepsilon}(t)\|_\infty \to 0 \text{ as } s \to \infty.$$

Note that

$$\sup_{t\in[0,1]} \|Fx^{v,\lambda,\varepsilon}(t) - Fp^{s,v,\lambda,\varepsilon}(t)\| \leq C \sup_{t\in[0,1]} \|x^{v,\lambda,\varepsilon}(t) - p^{s,v,\lambda,\varepsilon}(t)\|_\infty$$

where $C := \sup_{t,s\in[0,1]} \sum_{i=1}^n \|A_i(t)\| + \|D_i(s,t)\|$. By the same reasoning as before, for s high enough, the polynomial $p^{s,v,\lambda,\varepsilon}$ will be (strictly) feasible to our TV-SDP. Moreover,

$$\langle c, p^{s,v,\lambda,\varepsilon}\rangle_{\mathbf{L^n}} \to \langle c, x^{v,\lambda,\varepsilon}\rangle_{\mathbf{L^n}} \text{ as } s \to \infty.$$

To get the overall result, fix ε small enough, then λ small enough, then v large enough, and then s large enough. □

2.3.2 Finding the best polynomial solution to a TV-SDP via SDP

In this section, we show how one can find the best polynomial solution of a given degree to a TV-SDP. This is done by reformulating the problem as a semidefinite program. This formulation is based on the fact that any univariate polynomial matrix $X(t)$ that is positive semidefinite over an interval has a certain sum of squares representation of low degree. This representation can be found by semidefinite programming using the well-known connection (see e.g. [157, 122]) between sum of squares polynomials and SDPs.

Theorem 5. (See [67, Theorem 2.5], [153, Theorem 6.11], and see [27] for a history of related proofs) *Let $X \in \mathbf{S^m}_d$ be a univariate $m \times m$ polynomial matrix of degree d. If d is odd, then $X(t) \succeq 0$ $\forall t \in [0,1]$ if and only if there exist (not necessarily square) polynomial matrices B_1 and B_2 of degree $\frac{d-1}{2}$ such that*

$$X(t) = tB_1(t)^T B_1(t) + (1-t)B_2(t)^T B_2(t).$$

Similarly, if d is even, then $X(t) \succeq 0$ $\forall t \in [0,1]$ if and only if there exist (not necessarily square) polynomial matrices B_1 and B_2 of degree $\frac{d}{2}$ and $\frac{d}{2}-1$ respectively such that

$$X(t) = B_1(t)^T B_1(t) + t(1-t)B_2(t)^T B_2(t).$$

This theorem results in a semidefinite representation of polynomial matrices that are positive semidefinite on the interval $[0,1]$ as we describe next. This transformation is rather standard and can be traced back to the work of Nesterov [146].

Proposition 2. *Let d, m be positive integers. There exist two linear maps α_d^m (which maps $\mathcal{S}^{\frac{d+1}{2}m}$ to $\mathbf{S^m}$ when d is odd and $\mathcal{S}^{(\frac{d}{2}+1)m}$ to $\mathbf{S^m}$ when d is even) and β_d^m (which maps $\mathcal{S}^{\frac{d+1}{2}m}$ to $\mathbf{S^m}$ when d is odd and $\mathcal{S}^{\frac{d}{2}m}$ to $\mathbf{S^m}$ when d is even) such that for any $X \in \mathbf{S^m}_d$, $X(t) \succeq 0\ \forall t \in [0,1]$ if and only if one can find positive semidefinite matrices Q_1, Q_2 of appropriate sizes that satisfy the equation*

$$X = \alpha_d^m(Q_1) + \beta_d^m(Q_2).$$

Proof. Fix positive integers m and d. Let $Y \in \mathbf{S^m}_d$ be an $m \times m$ polynomial matrix of degree d. It is well known that Y can be written as $B(t)^T B(t)$ for some polynomial matrix B if and only if the polynomial $y^T Y(t) y$ is a sum of squares of some polynomials in the variables $(t, y_1, \ldots, y_m)$; see e.g. [118]. The latter condition is equivalent to existence of a $(\frac{d}{2}+1)m \times (\frac{d}{2}+1)m$ matrix $Q \succeq 0$ (d is necessarily even) such that the following polynomial identity holds

$$y^T Y(t) y = v(t,y)^T Q v(t,y), \tag{2.9}$$

where

$$v(t,y) = (y_1, \ldots, y_m, y_1 t, \ldots, y_m t, \ldots, y_1 t^{\frac{d}{2}}, \ldots, y_m t^{\frac{d}{2}})^T$$

is the vector of all monomials of the form $y_l t^k$ for $l = 1, \ldots, m$, and $k = 0, \ldots, \frac{d}{2}$; see e.g. [9, Section 3]. For notational convenience, we index the entries of the matrix Q by the monomials in $v := v(t,y)$. This means that when we write Q_{v_i, v_j}, we refer the (i,j)-th entry of Q.

Note that for any symmetric matrix Q, there exists a unique $Y \in \mathbf{S^m}$ that satisfies the identity (2.9). Indeed, considering the expression $v(t,y)^T Q v(t,y)$ as a polynomial in $y_1, \ldots, y_m$ with coefficients in $\mathbb{R}[t]$, the coefficient of $y_i y_j$ is equal to twice the (i,j)-th entry of $Y(t)$ when $i \neq j$, and equal to the (i,i)-the entry of $Y(t)$ otherwise. Define Λ_d^m to be the linear function that maps a symmetric matrix $Q \in \mathcal{S}^{(\frac{d}{2}+1)m}$ to the $m \times m$ polynomial matrix Y of degree d that satisfies identity (2.9), i.e.

$$Y = \Lambda_d^m(Q) \iff Y_{ij}(t) = c_{ij} \sum_{k,l \in \{0,\ldots,\frac{d}{2}\}} Q_{y_i t^k, y_j t^l} t^{k+l} \quad \forall i,j \in \{1,\ldots,m\}, \tag{2.10}$$

with $c_{ij} = \frac{1}{2}$ when $i \neq j$ and $c_{ii} = 1$.

We have just shown that an $m \times m$ polynomial matrix Y of degree d can be written as $Y(t) = B(t)^T B(t)$ for some polynomial matrix B if and only if there exists an $m(\frac{d}{2}+1) \times m(\frac{d}{2}+1)$ positive semidefinite matrix Q such that

$$Y = \Lambda_d^m(Q).$$

Combining this result with Theorem 5, we get that any $m \times m$ polynomial matrix X is positive semidefinite on $[0,1]$ if and only if there exist positive semidefinite matrices Q_1, Q_2 such that

$$X = \alpha_d^m(Q_1) + \beta_d^m(Q_2),$$

where

$$\begin{aligned} &\alpha_d^m(Q) = t\Lambda_{d-1}^m(Q) \text{ and } \beta_d^m(Q) = (1-t)\Lambda_{d-1}^m(Q) \text{ when } d \text{ is odd,} \\ &\alpha_d^m(Q) = \Lambda_d^m(Q) \text{ and } \beta_d^m(Q) = t(1-t)\Lambda_{d-2}^m(Q) \text{ when } d \text{ is even.} \end{aligned} \tag{2.11}$$

□

The next theorem summarizes the results of this subsection.

Theorem 6. *For $d \in \mathbb{N}$, the following SDP finds the best polynomial solution of degree d to the TV-SDP in (2.1) with data $c, A_0, A_1, \ldots, A_n, D_1, \ldots, D_n$:*

$$\begin{array}{ll} \max\limits_{x,Q_1,Q_2} & \langle c, x\rangle_{\mathbf{L^n}} \\ s.t. & x \in \mathbb{R}_d^n[t] \\ & Q_1, Q_2 \succeq 0, \\ & Fx = \alpha_{d'}^m(Q_1) + \beta_{d'}^m(Q_2). \end{array} \tag{2.12}$$

Here, d' is the degree of Fx, i.e.

$$d' = \max\{\deg(A_0), \max_{i=1,\ldots,n} \deg(A_i) + d, \max_{i=1,\ldots,n} \deg(D_i) + d + 1\}.$$

From a practical standpoint, a nice feature of this SDP hierarchy is the dimensions of the matrices Q_1, Q_2 grow only linearly with d.

Remark 2. *For implementation purposes, one does not need to explicitly write out the linear maps $\alpha_{d'}^m$ and $\beta_{d'}^m$ in Theorem 6. Certain solvers, such as YALMIP [131] or SOSTOOLS [150], accept the problem directly in the following form (and do the conversion to an SDP in the background):*

$$\begin{array}{ll} \max\limits_{x,X_1,X_2} & \langle c, x\rangle_{\mathbf{L^n}} \\ s.t. & x \in \mathbb{R}_d^n[t], X_1, X_2 \in \mathbf{S^m}_{d'-1}, \\ & Fx(t) = tX_1(t) + (1-t)X_2(t), \\ & y^T X_1(t) y, y^T X_2(t) y \text{ are sums of squares of polynomials,} \end{array}$$

when d is odd, and

$$\begin{array}{ll} \max\limits_{x,X_1,X_2} & \langle c, x\rangle_{\mathbf{L^n}} \\ s.t. & x \in \mathbb{R}_d^n[t], X_1 \in \mathbf{S^m}_{d'}, X_2 \in \mathbf{S^m}_{d'-2}, \\ & Fx(t) = X_1(t) + t(1-t)X_2(t), \\ & y^T X_1(t) y, y^T X_2(t) y \text{ are sums of squares of polynomials,} \end{array}$$

when d is even. The aforementioned linear maps however help us with the notation and presentation of the next section.

Remark 3. *The results of this section combined with Theorem 4 imply the following: For any strictly feasible TV-SDP, the SDPs in (2.12), indexed by the integer d, produce a sequence of* feasible *polynomials to the TV-SDP in (2.1), whose objective values converge to the optimal value of (2.1). Note that we are not making any claims about the convergence of the sequence of polynomials returned by these SDPs. Indeed, our interest is not for these polynomials to be close (in some distance measure) to an optimal solution of (2.1) (which might not even be continuous), but for them to be feasible and have arbitrarily good objective value.*

2.4 The Dual Approach: Obtaining Upper Bounds

In the previous section, we showed how one can obtain arbitrarily good lower bounds on the optimal value of a strictly feasible TV-SDP by searching for polynomial solutions of increasing degree. In practice, one often has a computational budget and cannot increase the degree of the candidate polynomials beyond a certain threshold. It is therefore very valuable to know how far the objective value of the best polynomial solution that one has found is from the optimal value of (2.1). Addressing this need is the subject of this section. More specifically, in Section 2.4.1 below we give a hierarchy of dual problems (or relaxations), indexed by a nonnegative integer d, that provide a sequence of improving upper bounds on the optimal value of (2.1). We show that under the boundedness assumption in (2.6), these upper bounds converge to the optimal value as $d \to \infty$. While the original formulation of the dual problems in Section 2.4.1 is infinite dimensional, we show in Section 2.4.2 that each of these problems can be solved exactly as an SDP of tractable size.

We remark that our dual problems are different to the best of our knowledge from those appearing in the literature on continuous linear programs. For instance, we can derive the Lagrangian dual of problem (2.1) using standard techniques from infinite-dimensional linear programming [16]. This problem would read

$$\begin{array}{ll} \inf\limits_{P \in \mathbf{S^m}^+([0,1])} & \langle P, A_0 \rangle_{\mathbf{S^m}} \\ \text{subject to} & \tilde{F}^* P(t) + c(t) = 0 \ \forall t \in [0,1] \text{ a.e.}, \end{array} \tag{2.13}$$

where $\tilde{F}^* P(t)$ is the vector of size n whose i-th component is given by the $(i+1)$-th component of $F^* P(t)$, with F^* as in (2.14) below. This is of course another TV-SDP, for which we can search for polynomial solutions to obtain upper bounds on the optimal value of (2.1). The reason we do not take this approach is that we do not want to make strict feasibility assumptions on both the primal and the dual (which our current proof strategy would require in order to ensure convergence of these bounds). Furthermore, more involved assumptions would likely be required to guarantee strong duality between the two TV-SDPs. Even in the special case of continuous linear programs, a number of assumptions are needed to obtain strong duality [86, 33].

The following definitions will be useful in the formulation of our dual problems.

Definition 2.4.1 (Adjoint maps)**.** *We define the adjoint of the affine map F in (2.2) to be the linear map $F^*\colon \mathbf{S^m} \to \mathbf{L^{n+1}}$ that acts on $P \in \mathbf{S^m}$ as follows:*

$$F^*P(t) = \begin{pmatrix} \operatorname{Tr}(A_0(t)P(t)) \\ \operatorname{Tr}(A_1(t)P(t)) + \int_t^1 \operatorname{Tr}(D_1(t,s)P(s))\,\mathrm{d}s \\ \vdots \\ \operatorname{Tr}(A_n(t)P(t)) + \int_t^1 \operatorname{Tr}(D_n(t,s)P(s))\,\mathrm{d}s \end{pmatrix}. \tag{2.14}$$

For an even positive integer d, the adjoint of the linear map Λ_d^m defined in (2.10) is the linear map $\Lambda_d^{m*}\colon \mathbf{S^m} \to \mathcal{S}^{(\frac{d}{2}+1)m}$ that acts on $P \in \mathbf{S^m}$ as follows:

$$\Lambda_d^{m*}(P) := \left(\int_0^1 t^{k+l} P_{ij}(t)\mathrm{d}t \right)_{y_i t^k, y_j t^l},$$

where the notation $Q_{y_i t^k, y_j t^l}$ stands for the (r,s) entry of the matrix $Q \in S^{(\frac{d}{2}+1)m}$ with r and s being the position of the monomials $y_i t^k$ and $y_j t^l$ in the vector

$$(y_1, \ldots, y_m, y_1 t, \ldots, y_m t, \ldots, y_1 t^{\frac{d}{2}}, \ldots, y_m t^{\frac{d}{2}})^T.$$

For $d \in \mathbb{N}$, the adjoints of the linear maps α_d^m and β_d^m defined in (2.11) are defined as follows:

$$\alpha_d^{m*}(P) := {\Lambda_{d-1}^m}^*(tP) \text{ and } \beta_d^{m*}(P) := {\Lambda_{d-1}^m}^*((1-t)P)\ \forall P \in \mathbf{S^m} \text{ when } d \text{ is odd},$$

$$\alpha_d^{m*}(P) := \Lambda_d^{m*}(P) \text{ and } \beta_d^{m*}(P) := {\Lambda_{d-2}^m}^*(t(1-t)P)\ \forall P \in \mathbf{S^m} \text{ when } d \text{ is even}.$$

The reader can check (using Fubini's double integration theorem when needed) that these adjoint maps satisfy the following identities.

Proposition 3. *For all $x \in \mathbf{L^n}$ and $P \in \mathbf{S^m}$,*

$$\langle Fx, P\rangle_{\mathbf{S^m}} = \langle \begin{pmatrix} 1 \\ x \end{pmatrix}, F^*P \rangle_{\mathbf{L^{n+1}}}.$$

For a linear map $L \in \{\Lambda_d^m, \alpha_d^m, \beta_d^m\}$, a polynomial matrix $P \in \mathbf{S^m}$, and a constant symmetric matrix Q of appropriate size,

$$\langle L(Q), P\rangle_{\mathbf{S^m}} = \operatorname{Tr}(QL^*(P)).$$

2.4.1 Dual formulation

To derive our dual problems, we start by observing that using Lemma 2.2.3, we can rewrite the TV-SDP in (2.1) as the following problem:

$$\begin{array}{ll} \max\limits_{x \in \mathbf{L^n}} & \langle c, x \rangle_{\mathbf{L^n}} \\ \text{subject to} & \langle Fx, P \rangle_{\mathbf{S^m}} \geq 0 \quad \forall P \in \mathbf{S^{m+}}([0,1]) \cap \mathbb{R}^{m \times m}[t]. \end{array}$$

To get an upper bound on the optimal value of (2.1), we relax the constraint in this problem by asking it to hold only for all $P \in \mathbf{S^{m+}}([0,1])_d$, i.e. for all polynomial matrices of degree bounded by some threshold d. This gives us our dual problem at level d, whose optimal value we denote by u_d:

$$\begin{array}{rll} u_d := \max\limits_{x \in \mathbf{L^n}} & \langle c, x \rangle_{\mathbf{L^n}} & \\ \text{subject to} & \langle Fx, P \rangle_{\mathbf{S^m}} \geq 0 & \forall P \in \mathbf{S^{m+}}([0,1])_d. \end{array} \tag{2.15}$$

Lemma 2.4.2 (Weak Duality)**.** *Let opt denote the optimal value of the TV-SDP in (2.1) and u_d be as in (2.15). Then, for all $d \in \mathbb{N}$, we have*

$$opt \leq u_d \text{ and } u_{d+1} \leq u_d.$$

Proof. Fix $d \in \mathbb{N}$. Since $\mathbf{S^{m+}}([0,1])_d \subseteq \mathbf{S^{m+}}([0,1])_{d+1}$, it is clear that $u_{d+1} \leq u_d$. Note that the only difference between problem (2.15) and the TV-SDP in (2.1) is that we have replaced the constraint $Fx \in \mathbf{S^{m+}}$ by $\langle Fx, P \rangle_{\mathbf{S^m}} \geq 0 \quad \forall P \in \mathbf{S^{m+}}([0,1])_d$. By Lemma 2.2.3, the former constraint is stronger than the latter. Therefore, opt $\leq u_d$. □

To get strong duality, we will make the additional assumption in (2.6); i.e. we assume that there exists a positive scalar γ such that

$$Fx \in \mathbf{S^{m+}} \implies \|x(t)\|_\infty < \gamma \ \forall t \in [0,1] \text{ a.e..}$$

We further require that this constraint already be included in F. In other words, F is taken to be of the form

$$Fx := \begin{pmatrix} \hat{F}x & 0 & 0 \\ 0 & \gamma I - diag(x) & 0 \\ 0 & 0 & diag(x) - \gamma I \end{pmatrix}, \tag{2.16}$$

where $\hat{F}x \succeq 0$ denotes the remaining constraints of the TV-SDP.

Theorem 7 (Strong Duality)**.** *Suppose that the TV-SDP in (2.1) satisfies the boundedness assumption (2.6) as explicitly imposed by a map F of the form (2.16). Let $opt \in \mathbb{R} \cup \{-\infty\}$ denote the optimal value of this TV-SDP. Then the optimal value u_d of problem (2.15) converges to opt as $d \to \infty$.*

Proof. From Lemma 2.4.2, the sequence $\{u_d\}$ is nonincreasing and bounded below by opt. It therefore converges to a (possibly infinite) limit $u^* \geq \text{opt}$. To conclude the proof, we show that $u^* \leq \text{opt}$.

Observe first that if there exists a nonnegative integer d such that $u_d = -\infty$, then $u^* = \text{opt} = -\infty$ and we are done. We can therefore suppose that the sequence $\{u_d\}$ never takes the value $-\infty$. We claim that when $d \geq \deg(c)$, u_d cannot take the value $+\infty$ either. To see why, fix $d \geq \deg(c)$ and let $x \in \mathbf{L^n}$ be any function that satisfies $\langle Fx, P\rangle_{\mathbf{S^m}} \geq 0 \; \forall P \in \mathbf{S^{m+}}([0,1])_d$. For any $q \in \mathbb{R}^n_d[t]$ that is elementwise nonnegative on $[0,1]$, by taking

$$P \in \left\{ \begin{pmatrix} 0 & 0 & 0 \\ 0 & diag(q) & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & diag(q) \end{pmatrix} \right\},$$

we get that $|\langle q, x\rangle_{\mathbf{L^n}}| \leq \gamma \langle q, \mathbf{1}\rangle_{\mathbf{L^n}}$. Let

$$m_c := \min_{i=1,\ldots,n} \min_{t\in[0,1]} c_i(t)$$

and observe that the polynomial $p(t) := c(t) - m_c\mathbf{1}$ is elementwise nonnegative on $[0,1]$. We therefore have

$$\begin{aligned} |\langle c, x\rangle_{\mathbf{L^n}}| &\leq |\langle c - m_c\mathbf{1}, x\rangle_{\mathbf{L^n}}| + |\langle m_c\mathbf{1}, x\rangle_{\mathbf{L^n}}| \\ &= |\langle p, x\rangle_{\mathbf{L^n}}| + |m_c||\langle \mathbf{1}, x\rangle_{\mathbf{L^n}}| \\ &\leq \gamma\langle p, \mathbf{1}\rangle_{\mathbf{L^n}} + |m_c|\gamma\langle \mathbf{1}, \mathbf{1}\rangle_{\mathbf{L^n}}, \end{aligned}$$

which proves our claim that u_d is finite.

As a consequence, for any integer $d \geq \deg(c)$ and for any positive scalar ε, there exists a function $x^{\varepsilon,d} \in \mathbf{L^n}$ such that

$$\langle Fx^{\varepsilon,d}, P\rangle_{\mathbf{S^m}} \geq 0 \; \forall P \in \mathbf{S^{m+}}([0,1])_d \text{ and } \langle c, x^{\varepsilon,d}\rangle_{\mathbf{L^n}} \geq u_d - \varepsilon. \tag{2.17}$$

For a given $\varepsilon > 0$, fix such a sequence $\{x^{\varepsilon,d}\}$ indexed by d. We claim that $\{x^{\varepsilon,d}\}$ must have a subsequence that converges weakly to a fucntion $x^\varepsilon \in \mathbf{L^n}$. Indeed, if we let $\mu^{\varepsilon,d} := \frac{l_{x^{\varepsilon,d}}}{\gamma}$, then for any polynomial $p \in \mathbb{R}^n_d[t]$ that is elementwise nonnegative on $[0,1]$, we have $|\mu^{\varepsilon,d}(p)| \leq \sum_{i=1}^n \int_0^1 p_i(t)\, dt$. Our claim then follows from Theorem 1. It is clear by weak convergence that $\langle c, x^\varepsilon\rangle_{\mathbf{L^n}} \geq u^* - \varepsilon$. Moreover, for any $P \in \mathbf{S^{m+}}([0,1]) \cap \mathbb{R}^{m\times m}[t]$, if $d \geq \deg(P)$, we have $\langle Fx^{\varepsilon,d}, P\rangle_{\mathbf{S^m}} \geq 0$, or equivalently

$$\left\langle \begin{pmatrix} 1 \\ x^{\varepsilon,d} \end{pmatrix}, F^*P \right\rangle_{\mathbf{L^{n+1}}} \geq 0.$$

Hence, by taking $d \to \infty$, $\left\langle \begin{pmatrix} 1 \\ x^{\varepsilon} \end{pmatrix}, F^*P \right\rangle_{\mathbf{L^{n+1}}} \geq 0$, showing that $\langle Fx^\varepsilon, P\rangle_{\mathbf{S^m}} \geq 0$. By Lemma 2.2.3, we conclude that $Fx^\varepsilon \in \mathbf{S^{m+}}([0,1])$ and therefore $\langle c, x^\varepsilon\rangle_{\mathbf{L^n}} \leq \text{opt}$.

We have just proven that for any $\varepsilon > 0$, there exists a feasible solution x^ε to the TV-SDP in (2.1) such that

$$u^* - \varepsilon \leq \langle c, x^\varepsilon \rangle_{\mathbf{L^n}} \leq \text{opt}.$$

This means that $u^* \leq \text{opt}$. □

2.4.2 The dual problem is an SDP

In this section, we show that the infinite-dimensional problem in (2.15) can be converted to an SDP of tractable size.

Theorem 8. *Consider problem (2.15) at level d and with data $c, A_0, \ldots, A_n, D_1, \ldots, D_n$. Let*

$$\hat{d} := \max\{\deg(c), \max_{i=1,\ldots,n} d + \deg(A_i), \max_{i=1,\ldots,n} d + 1 + \deg(D_i)\}.$$

The optimal value of problem (2.15) does not change when the space $\mathbf{L^n}$ is replaced with $\mathbb{R}^n_{\hat{d}}[t]$. Moreover, this optimal value is equal to the optimal value of the following SDP

$$\begin{array}{ll} \max\limits_{x \in \mathbb{R}^n_{\hat{d}}[t]} & \langle c, x \rangle_{\mathbf{L^n}} \\ \textit{subject to} & \alpha_d^{m*}(Fx) \succeq 0 \\ & \beta_d^{m*}(Fx) \succeq 0, \end{array} \tag{2.18}$$

where the adjoint maps $\alpha_d^{m*}, \beta_d^{m*}$ are as in Definition 2.4.1.

Just like our primal hierarchy, observe that the dimensions of the matrices that need to be positive semidefinite in this SDP hierarchy grow only linearly with d. We start with a simple and standard lemma that will help us prove the first claim of the theorem.

Lemma 2.4.3. *For any function $f \in \mathbf{L^1}$, there exists a sequence of polynomials $\{p_d\}$ such that for every $d \in \mathbb{N}$, the polynomial p_d has degree d and satisfies*

$$\int_0^1 q(t) p_d(t)\, \mathrm{d}t = \int_0^1 q(t) f(t)\, \mathrm{d}t \quad \forall q \in \mathbb{R}_d[t]. \tag{2.19}$$

Proof. Fix $d \in \mathbb{N}$. Parameterize a generic univariate polynomial $p(t)$ of degree d as

$$p(t) = \sum_{i=0}^{d} p_i t^i$$

and let $m_i := \int_0^1 t^i f(t)\, \mathrm{d}t$ for $i = 0, \ldots, d$. By linearity, the equality in (2.19) is equivalent to

$$\int_0^1 t^i p(t)\, \mathrm{d}t = m_i \quad i = 0, \ldots, d.$$

Let H denote the $(d+1)\times(d+1)$ matrix whose (i,j)-th entry is equal to $\int_0^1 t^{i+j-2}\,\mathrm{d}t = \frac{1}{i+j-1}$. Equation (2.19) is therefore equivalent to

$$\begin{pmatrix} p_0, & \dots & , p_d \end{pmatrix} H = \begin{pmatrix} m_0, & \dots, & m_d \end{pmatrix}.$$

It follows that this equation has a (unique) solution as the matrix H (often named the *Hilbert* matrix) is known to be invertible [98]. □

Proof. of Theorem 8 Fix $d \in \mathbb{N}$. Let $x \in \mathbf{L^n}$ be a feasible solution to (2.15), i.e. satisfy

$$\langle Fx, P\rangle_{\mathbf{S^m}} \geq 0 \; \forall P \in \mathbf{S^{m+}}([0,1])_d. \tag{2.20}$$

Notice that this expression depends on x only through its $\hat{d}$ moments. More precisely, if a function $y \in \mathbf{L^n}$ satisfies

$$\langle q, x\rangle_{\mathbf{L^n}} = \langle q, y\rangle_{\mathbf{L^n}} \; \forall q \in \mathbb{R}^n_{\hat{d}}[t], \tag{2.21}$$

then for all $P \in \mathbf{S^{m+}}([0,1])_d$,

$$\langle Fy, P\rangle_{\mathbf{S^m}} = \langle \begin{pmatrix} 1 \\ y \end{pmatrix}, F^*P\rangle_{\mathbf{L^{n+1}}} = \langle \begin{pmatrix} 1 \\ x \end{pmatrix}, F^*P\rangle_{\mathbf{L^{n+1}}} = \langle Fx, P\rangle_{\mathbf{S^m}} \geq 0.$$

Furthermore, $\langle c, y\rangle_{\mathbf{L^n}} = \langle c, x\rangle_{\mathbf{L^n}}$. By Lemma 2.4.3, there always exists a function y in $\mathbb{R}^n_{\hat{d}}[t]$ that satisfies (2.21). Therefore, we can restrict the space $\mathbf{L^n}$ in problem (2.15) to $\mathbb{R}^n_{\hat{d}}[t]$. Now if $x \in \mathbb{R}^n_{\hat{d}}[t]$, by Proposition 2, condition (2.20) is equivalent to

$$\langle Fx, \alpha^m_d(Q_1) + \beta^m_d(Q_2)\rangle_{\mathbf{S^m}} \geq 0 \; \forall Q_1 \succeq 0, \forall Q_2 \succeq 0,$$

which itself is equivalent to

$$\langle Fx, \alpha^m_d(Q_1)\rangle_{\mathbf{S^m}} \geq 0 \; \forall Q_1 \succeq 0, \text{ and } \langle Fx, \beta^m_d(Q_2)\rangle_{\mathbf{S^m}} \geq 0 \; \forall Q_2 \succeq 0.$$

By Proposition 3, this latter statement holds if and only if

$$\langle \alpha^{m*}_d(Fx), Q_1\rangle \geq 0 \; \forall Q_1 \succeq 0, \text{ and } \langle \beta^{m*}_d(Fx), Q_2\rangle \geq 0 \; \forall Q_2 \succeq 0,$$

i.e.,

$$\alpha^{m*}_d(Fx) \succeq 0 \text{ and } \beta^{m*}_d(Fx) \succeq 0.$$

□

2.5 Applications

In this section, we present three applications of time-varying semidefinite programs along with some numerical experiments.

2.5.1 Time-varying Max-Flow

In our first example, we study a generalization of the classical maximum-flow problem where the pipeline capacities are allowed to vary with time. More specifically, we are given a graph with node set $V := \{1, \dots n\}$, and edge set $E \subseteq [n]^2$. We take node 1 to be the source of the flow and node n to be the target. Our decision variables are functions $f_{ij} \in \mathbf{L^1}$, for $(i,j) \in E$, with $f_{ij}(t)$ denoting the instantaneous flow on edge (i,j) at time $t \in [0,1]$. We have as input functions $b_{i,j} \in \mathbb{R}[t]$ with $b_{ij}(t)$ denoting the capacity of edge (i,j) at time $t \in [0,1]$.

The capacity (and nonnegativity) constraints that we need to satisfy are

$$0 \leq f_{ij}(t) \leq b_{ij}(t) \quad \forall (i,j) \in E,\ \forall t \in [0,1] \text{ a.e..}$$

We further need to satisfy conservation of flow constraints at every node other than the source and the target nodes:

$$\sum_{j:(i,j)\in E} f_{ij}(t) - \sum_{j:(j,i)\in E} f_{ji}(t) = 0 \quad \forall i \in V \setminus \{1,n\},\ \forall t \in [0,1] \text{ a.e..}$$

In some applications, a subset of the edges that we denote by $E_1 \subseteq E$ may not be able to handle an instantaneous change in the flow that is too large. In other words, we need to impose the following additional constraints:

$$\left| \frac{\mathrm{d}}{\mathrm{d}t} f_{ij}(t) \right| \leq b_{ij}^{\text{deriv}}(t) \quad \forall (i,j) \in E_1,\ \forall t \in [0,1] \text{ a.e.,} \tag{2.22}$$

for some pre-specified functions $b_{ij}^{\text{deriv}} \in \mathbb{R}[t]$. We handle this by introducing a new decision variable $g_{ij} \in \mathbf{L^1}$ for every $(i,j) \in E_1$ and imposing

$$\int_0^t g_{ij}(s)\ \mathrm{d}s - f_{ij}(t) = 0 \text{ and } -b_{ij}^{\text{deriv}}(t) \leq g_{ij}(t) \leq b_{ij}^{\text{deriv}}(t) \quad \forall (i,j) \in E_1,\ \forall t \in [0,1] \text{ a.e..}$$

Moreover, we assume that because of limitations on production of the flow at the source node, the cumulative flow going into the network up to time t cannot exceed $b^{\text{cum}}(t)$ for some given function $b^{\text{cum}} \in \mathbb{R}[t]$. Hence, this constraint reads

$$\int_0^t \sum_{(1,j)\in E} f_{1j}(t)dt \leq b^{\text{cum}}(t) \quad \forall t \in [0,1] \text{ a.e..} \tag{2.23}$$

Our objective is to send as much flow as possible from the source to the target node over the time interval $[0,1]$. Hence, the overall problem, which is a time-varying semidefinite (in fact, linear) program, reads:

$$
\begin{aligned}
\max_{f_{ij},g_{ij}} \quad & \int_0^1 \sum_{(1,j)\in E} f_{1j}(t)\ \mathrm{d}t \\
& \left.\begin{array}{ll}
0 \le f_{ij}(t) \le b_{ij}(t) & \forall (i,j) \in E \\
\displaystyle\sum_{j:(i,j)\in E} f_{ij}(t) - \sum_{j:(j,i)\in E} f_{ji}(t) = 0 & \forall i \in V \setminus \{1,n\} \\
\int_0^t g_{ij}(s)\ \mathrm{d}s - f_{ij}(t) = 0 & \forall (i,j) \in E_1 \\
-b^{\mathrm{deriv}}(t) \le g_{ij}(t) \le b^{\mathrm{deriv}}(t) & \forall (i,j) \in E_1 \\
\displaystyle\int_0^t \sum_{(1,j)\in E} f_{1j}(s)\ \mathrm{d}s \le b^{\mathrm{cum}}(t) &
\end{array}\right\} \forall t \in [0,1] \text{ a.e..}
\end{aligned}
\tag{2.24}
$$

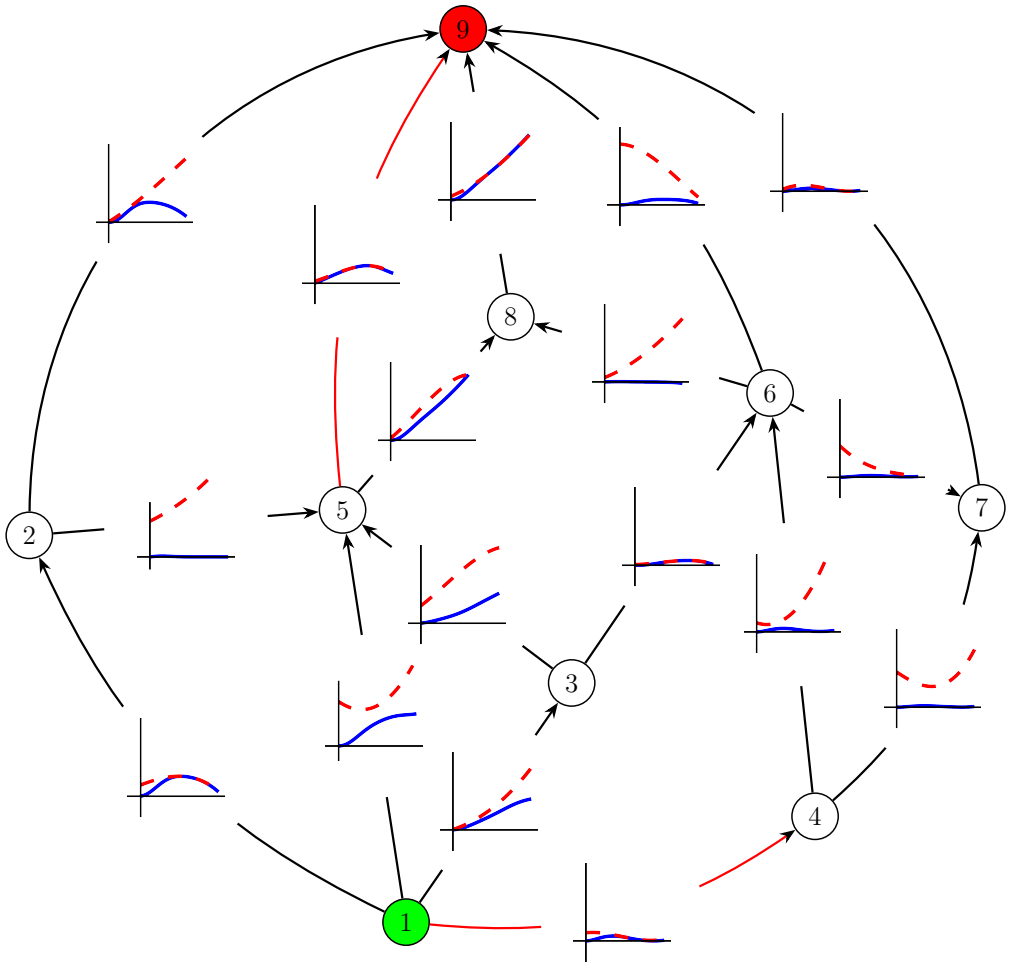


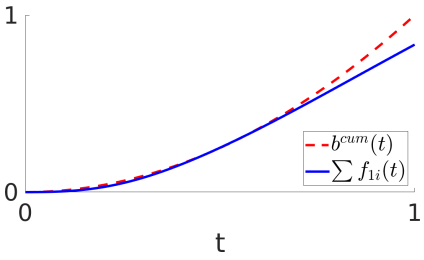
Figure 2.2: An instance of the time-varying max-flow problem. The edge capacities $b_{ij}(t)$, over the time interval $[0,1]$, are plotted with red dotted lines. The optimal polynomial flow $f_{ij}(t)$ of degree at most 10 is plotted on each edge with solid blue lines.

As a numerical example, we consider the network in Figure 2.2 with capacities $b_{ij}(t)$ plotted with red dotted lines on each edge (i, j). Each of these polynomials b_{ij} is a nonnegative polynomial of degree 3 that is generated as follows

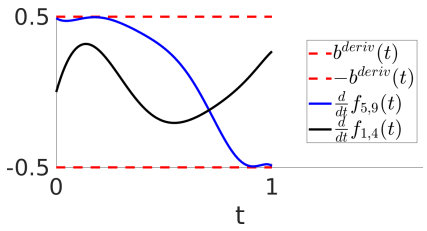
$$b_{ij}(t) = t(a_{ij}^{(1)} + a_{ij}^{(2)}\ t)^2 + (1-t)(a_{ij}^{(3)} + a_{ij}^{(4)}\ t)^2, \tag{2.25}$$

where $a_{ij}^{(k)}$ are generated independently and uniformly at random from $[-1, 1]$. We take $E_1 = \{(1, 4), (5, 9)\}$, $b^{\text{deriv}}(t) = \frac{1}{2}$, and $b^{\text{cum}}(t) = t^2$.

Using the machinery of Section 2.3, we solve semidefinite programs (as given in Theorem 6) that find the best polynomial solution of degree $d \in \{2, 3, \ldots, 10\}$ to the TV-SDP in (2.24). The optimal values of these problems, which provide improving lower bounds on the optimal value of problem (2.24), are reported in the first row of Table 2.1. We also plot the best polynomial solution of degree 10 on each edge of the graph in Figure 2.2 with solid blue lines. Figure 2.3 shows that this solution satisfies the constraints in (2.22) and (2.23).



(a) The cumulative flow $\sum_{(1,i)\in E} f_{1i}(t)$ at time t going through the network and the maximum flow available $b^{\text{cum}}(t)$ up to that time.



(b) The derivative of the flow going through the edges in E_1 and the maximum rate of change $b^{\text{deriv}}(t)$ and $-b^{\text{deriv}}(t)$ allowed for the flow.

Figure 2.3: Plots demonstrating that the constraints in (2.22) and (2.23) are satisfied by the best polynomial solution of degree 10 for (2.24).

We also use the machinery of Section 2.4 to solve the dual problems in (2.15) in order to get upper bounds on the optimal value of the TV-SDP in (2.24). By Theorem 8, the dual problem at level d is equivalent (after some rewriting) to the following

SDP:

$$
\begin{array}{ll}
\max\limits_{f_{ij},g_{ij}\in\mathbb{R}_{d+1}[t]} & \int_0^1 \sum\limits_{(1,j)\in E} f_{1j}(t)dt \\
& \alpha_d^{1*}(b_{ij}-f_{ij}) \succeq 0, \qquad \beta_d^{1*}(b_{ij}-f_{ij}) \succeq 0 \qquad \forall (i,j)\in E \\
& \alpha_d^{1*}(f_{ij}) \succeq 0, \qquad \beta_d^{1*} f_{ij} \succeq 0 \qquad \forall (i,j)\in E \\
& \alpha_d^{1*}\left(\sum\limits_{j:(i,j)\in E} f_{ij} - \sum\limits_{j:(j,i)\in E} f_{ji}\right) = 0, \quad \beta_d^{1*}\left(\sum\limits_{j:(i,j)\in E} f_{ij} - \sum\limits_{j:(j,i)\in E} f_{ji}\right) = 0 \quad \forall i \in V\setminus\{1,n\} \\
& \alpha_d^{1*}\left(\int_0^t g_{ij}(s)\,\mathrm{d}s - f_{ij}\right) = 0, \qquad \beta_d^{1*}\left(\int_0^t g_{ij}(s)\,\mathrm{d}s - f_{ij}\right) = 0 \qquad \forall (i,j)\in E_1 \\
& \alpha_d^{1*}\left(b^{\mathrm{deriv}} - g_{ij}\right) \succeq 0, \qquad \beta_d^{1*}\left(b^{\mathrm{deriv}} - g_{ij}\right) \succeq 0 \qquad \forall (i,j)\in E_1 \\
& \alpha_d^{1*}\left(b^{\mathrm{deriv}} + g_{ij}\right) \succeq 0, \qquad \beta_d^{1*}\left(b^{\mathrm{deriv}} + g_{ij}\right) \succeq 0 \qquad \forall (i,j)\in E_1 \\
& \alpha_d^{1*}\left(b^{\mathrm{cum}} - \int_0^t \sum\limits_{(1,j)\in E} f_{1j}(s)\right) \succeq 0, \quad \beta_d^{1*}\left(b^{\mathrm{cum}} - \int_0^t \sum\limits_{(1,j)\in E} f_{1j}(s)\right) \succeq 0.
\end{array}
\tag{2.26}
$$

The optimal value of this problem for different values of d is reported in the second row of Table 2.1.

d	2	3	4	5	6	7	8	9	10
lower bound	0.7201	0.7952	0.8170	0.8267	0.8274	0.8277	0.8279	0.8281	0.8282
upper bound	0.8700	0.8574	0.8541	0.8455	0.8446	0.8431	0.8421	0.8419	0.8413

Table 2.1: Upper and lower bounds on the optimal value of the time-varying max-flow problem in (2.24). In the first row, we report the objective value of the best polynomial solution of degree d. In the second row, we report the optimal value of the dual problem in (2.15) at level d.

Note from the two tables that the objective value of the degree-10 polynomial solution we have found is guaranteed to be within 2% of the best objective value possible. The running time of our largest SDPs on a standard laptop with the solver MOSEK [22] is in the order of a second. If we increase the degree much beyond 10, our solver runs into numerical issues. This is not surprising as we are formulating our SDPs using the standard monomial basis. Much improvement is possible on the implementation front using e.g. the ideas in [132, 153, 152, 154]. Such implementation improvements are left for future work.

2.5.2 A time-varying wireless coverage problem

In our second example, we present an application to wireless coverage of a targeted geographical region which moves over time. This is a time-varying generalization of problems considered in [58, 59, 60, 7]. In this setting, we have n_T wireless electromagnetic transmitters located at known locations $\bar{T}_i = (\bar{x}_i, \bar{y}_i)$ on the plane. Each transmitter $i \in \{1,\ldots,n_T\}$ is an omnidirectional power source providing a signal strength of $E_i(t,x,y)$ at time t in location (x,y) on the plane. Laws of electromag-

netic wave propagation stipulate that

$$E_i(x,y,t) = \frac{c_i(t)}{(x-\bar{x}_i)^2 + (y-\bar{y}_i)^2},$$

where $c_i(t)$, which is our decision variable, is the transmission power of the transmitter i at time t. There are n_R regions on the plane that move over time and that need to be covered with sufficient signal strength. For $j \in \{1,\ldots,n_R\}$ and $t \in [0,1]$, we define each such region $\mathcal{B}_j(t)$ with k_j polynomial inequalities:

$$\mathcal{B}_j(t) := \{(x,y) \in \mathbb{R}^2 | \quad g_{t,j,k}(x,y) \geq 0, k = 1,\ldots,k_j\}.$$

Here, for $j = 1,\ldots,n_R$, $k = 1,\ldots,k_j$, $g_{t,j,k}(x,y)$ is a polynomial in (x,y) whose coefficients depend on t. We further assume that for $j = 1,\ldots,n_R$ and for all $t \in [0,1]$,

$$g_{t,j,1}(x,y) = r^2 - x^2 - y^2$$

for some large enough scalar r.

Our goal is to ensure that for all time $t \in [0,1]$, the strength of the signal in all regions $\mathcal{B}_j(t)$ is at least a given threshold C. In other words, our constraints in this problem are

$$\left.\begin{array}{rl} E(x,y,t) := \sum_{i=1}^{n_T} E_i(x,y,t) \geq C & \forall (x,y) \in \mathcal{B}_j(t), \forall j \in \{1,\ldots,n_R\} \\ c_i(t) \geq 0 & \forall i \in \{1,\ldots,n_T\} \end{array}\right\} \quad \forall t \in [0,1] \text{ a.e..} \tag{2.27}$$

Our objective is to minimize the total cost of power generation, which is directly proportional to

$$\int_0^1 \sum_{i=1}^{n_T} c_i(t)\, \mathrm{d}t.$$

Notice that the first inequality in (2.27) is an inequality involving rational functions. Upon taking common denominators, we can reformulate this constraint as

$$p_t(x,y) := -C \prod_{i=1}^{n_T} [(x-\bar{x}_i)^2 + (y-\bar{y}_i)^2] + \sum_{i=1}^{n_T} c_i(t) \prod_{k \neq i} [(x-\bar{x}_k)^2 + (y-\bar{y}_k)^2] \geq 0$$
$$\forall (x,y) \in \mathcal{B}_j(t), \forall j = 1,\ldots,n_R, \forall t \in [0,1] \text{ a.e..} \tag{2.28}$$

Note that $p_t(x,y)$ is a polynomial in (x,y) whose coefficients depend on t. Let $v_{\tilde{d}}$ denote the vector of monomials in (x,y) of degree up to $\tilde{d}$, i.e.

$$v_{\tilde{d}} := v_{\tilde{d}}(x,y) = (1, x, \ldots, x^{\tilde{d}}, xy, \ldots, x^{\tilde{d}-1}y, \ldots, y^{\tilde{d}})^T.$$

d	2	3	4	5	6	7	8	9	10
	$+\infty$	56.64	54.52	54.43	54.14	54.14	53.95	53.94	53.93

Table 2.2: Objective values of optimal polynomial solutions of degree d to the time-varying wireless coverage problem in (2.30).

It is easy to check that for fixed $j \in \{1, \ldots, n_R\}, t \in [0,1]$, existence of positive semidefinite matrices $P_0^{(j)}(t), \ldots, P_{k_j}^{(j)}(t)$ satisfying the polynomial identity

$$p_t(x,y) = v_{\tilde{d}}(x,y)^T P_0^{(j)}(t) v_{\tilde{d}}(x,y) + \sum_{k=1}^{k_j} v_{\tilde{d}}(x,y)^T P_k^{(j)}(t) v_{\tilde{d}}(x,y) g_{t,j,k}(x,y) \qquad (2.29)$$

implies the constraint in (2.28). Conversely, for every fixed $j \in \{1, \ldots, n_R\}$ and $t \in [0,1]$, Putinar's Positivstellensatz [39] implies that if the constraint in (2.28) is satisfied strictly, one can always find a nonnegative integer $\tilde{d}$ and matrices $P_0^{(j)}(t), \ldots, P_{k_j}^{(j)}(t)$ that satisfy (2.29).

For any fixed $\tilde{d} \in \mathbb{N}$, our overall problem is the following TV-SDP:

$$\begin{array}{ll} \min\limits_{c_i, P_k^{(j)}} & \int_0^1 \sum_{i=1}^{n_T} c_i(t)\, \mathrm{d}t \\ & c_i \in \mathbf{L^1} \quad i = 1, \ldots, n_T \\ & P_k^{(j)} \in \mathbf{S}^{\frac{(\tilde{d}+1)(\tilde{d}+2)}{2}} \quad k = 0, \ldots, k_j,\ j = 1, \ldots, n_R \\ & \left.\begin{array}{ll} c_i(t) \geq 0 & i = 1, \ldots, n_T \\ p_t(x,y) = v_{\tilde{d}}^T P_0^{(j)}(t) v_{\tilde{d}} + \sum_{k=1}^{k_j} g_{t,j,k}(x,y) v_{\tilde{d}}^T P_k^{(j)}(t) v_{\tilde{d}} & \forall (x,y) \in \mathbb{R}^2,\ j = 1, \ldots, n_R \\ P_k^{(j)}(t) \succeq 0 & k = 0, \ldots, k_j, j = 1, \ldots, n_R \end{array}\right\} \forall t \in [0,1] \text{ a.e..} \end{array} \qquad (2.30)$$

Note that constraint (2.28) that appears in the TV-SDP in (2.30) is an equality between two polynomials in (x,y). Since two polynomials are equal if and only if their coefficients match, this constraint can be rewritten as a finite number of linear equations in our decision variables.

We now solve a numerical example with the following data:

$$C = 1, n_T = 2, \bar{T}_1 = (0,0), \bar{T}_2 = (5,5), n_R = 2, k_1 = k_2 = 2,$$

$$r = 10,\ g_{t,1,2}(x,y) = 1 - \left((x - 3t + 3)^2 + (y - 5t)^2\right),\ g_{t,2,2}(x,y) = 1 - \left(x^2 + (y - 5t + 1))^2\right).$$

In other words, our two regions are disks of unit radius whose centers move with time.

We start by finding polynomial solutions $c_1, c_2 \in \mathbb{R}_d[t]$ that satisfy the nonnegativity and the signal strength requirements in (2.27). For this, we solve the TV-SDP in (2.30) with $\tilde{d} = 1$. Using the methodology of Section 2.3.2, we solve semidefinite programs (as given in Theorem 6) to obtain the best polynomial solution of degree $d \in \{2, 3, \ldots, 10\}$. The objective values of the optimal solutions are reported in Table 2.2.

Note that if we do not allow the solution to depend on time (or even if we allow it to depend on time as a polynomial of degree less than 3), then the TV-SDP in (2.30) becomes infeasible. As we increase the degree, the problem becomes feasible and the objective value improves.

Figure 2.4 demonstrates a sanity check on our solution at six snapshots of time. Indeed, the two regions $\mathcal{B}_1(t)$ and $\mathcal{B}_2(t)$ are receiving a signal of strength of at least 1.

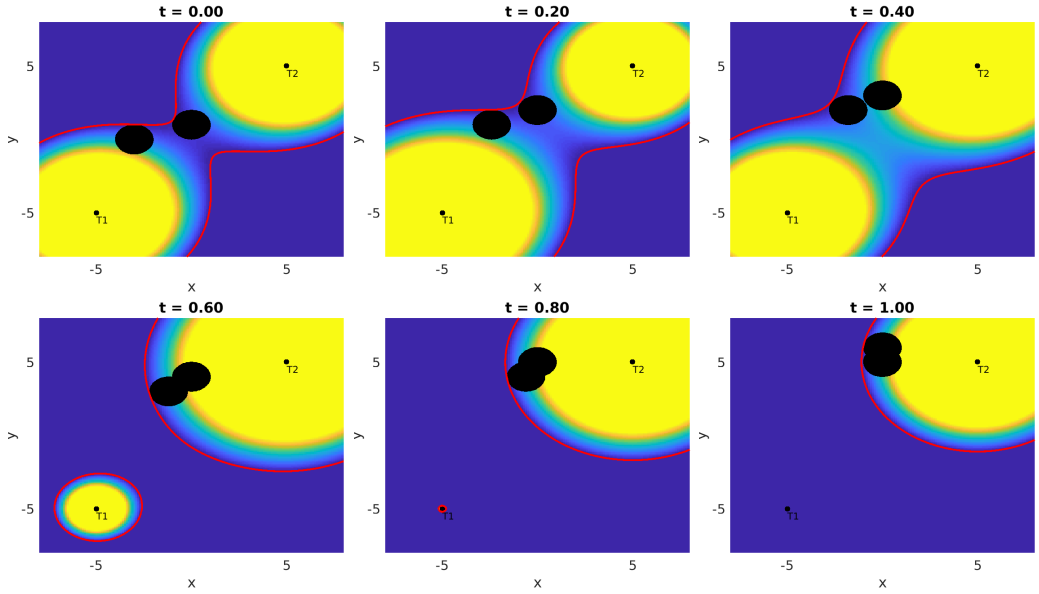


Figure 2.4: Six time snapshots—at $t = 0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1$—of the wireless coverage obtained by the best polynomial solution of degree 10. The two time-varying regions $\mathcal{B}_1(t)$ and $\mathcal{B}_2(t)$ that need to receive a signal strength of at least 1 at all times $t \in [0, 1]$ are colored in black. The heatmap in the background demonstrates the signal strength at each location with light yellow representing high and dark blue representing low signal strengths. The region delimited by the red curves is guaranteed to receive a signal strength of at least 1.

To have an idea of how far our best polynomial solution of degree 10 is from being optimal to the TV-SDP in (2.30), we solve the dual problem (2.15) presented in Section 2.4. After some rewriting, this dual problem at level d becomes the following SDP:

$$
\begin{aligned}
&\min_{c_i, P_k^{(j)} \text{ of degree } d+2} \int_0^1 \textstyle\sum_{i=1}^{n_T} c_i(t)\, \mathrm{d}t \\
&\text{subject to} \\
&\quad \alpha_d^{1*}(c_i) \succeq 0, \beta_d^{1*}(c_i) \succeq 0 && i = 1, \ldots, n_T \\
&\quad \alpha_d^{1*}\left(p_t(x,y) - v_{\tilde{d}}^T P_0^{(j)}(t) v_{\tilde{d}} + \textstyle\sum_{k=1}^{k_j} g_{t,j,k}(x,y) v_{\tilde{d}}^T P_k^{(j)}(t) v_{\tilde{d}}\right) = 0 && \forall (x,y) \in \mathbb{R}^2,\ j = 1, \ldots, n_R \\
&\quad \beta_d^{1*}\left(p_t(x,y) - v_{\tilde{d}}^T P_0^{(j)}(t) v_{\tilde{d}} + \textstyle\sum_{k=1}^{k_j} g_{t,j,k}(x,y) v_{\tilde{d}}^T P_k^{(j)}(t) v_{\tilde{d}}\right) = 0 && \forall (x,y) \in \mathbb{R}^2,\ j = 1, \ldots, n_R \\
&\quad \alpha_d^{m*}(P_k^{(j)}) \succeq 0, \beta_d^{m*}(P_k^{(j)}) \succeq 0 && k = 1, \ldots, n_T,\ j = 1, \ldots, n_R,
\end{aligned}
\tag{2.31}
$$

where $m = \frac{(\tilde{d}+1)(\tilde{d}+2)}{2}$. Note that the second and third set of constraints are requiring a polynomial matrix in (x, y) whose coefficients depend linearly on the decision variables to be identically zero. Once again, this is simply a finite numbers of equality constraints.

The optimal value of problem (2.31) with $d = 10$ is equal to 52.66. This tells us that the objective value of the degree-10 polynomial solution reported in Table 2.2 is within 2.5% of the optimal value of the TV-SDP in (2.30).

2.5.3 Bi-objective SDP and Pareto curve approximation

In our third and last example, we formulate a bi-objective (non time-varying) semidefinite program as a time-varying SDP.

A bi-objective semidefinite program is a standard SDP that involves two objective functions. More precisely, we are concerned with the simultaneous maximization of two objective functions

$$\langle c_1, x\rangle \text{ and } \langle c_2, x\rangle$$

over the feasible set

$$\mathcal{F} := \{x \in \mathbb{R}^n \mid Fx := A_0 + \sum_{i=1}^{n} x_i A_i \succeq 0\},$$

where $A_0, \ldots, A_n$ are given by $m \times m$ symmetric matrices. In general there exists no single solution x that maximizes both objective functions at the same time. As a trade-off, one is interested in solving the following problem

$$y(t) := \begin{array}{ll} \max\limits_{x\in\mathbb{R}^n} & \langle c_1, x\rangle \\ \text{subject to} & \langle c_2, x\rangle \geq t \text{ and } Fx \succeq 0, \end{array} \tag{2.32}$$

for various values of t. In the case where $\mathcal{F}$ is compact, we can without loss of generality take t to vary in $[0, 1]$ after a possible rescaling. This gives rise to the following trade-off curve, which we refer to as the Pareto curve:

$$PC := \{(t, y(t)) \mid t \in [0, 1]\}.$$

Any point on this curve tells us that in order to improve the first objective function beyond $y(t)$, the second objective needs to necessarily be smaller than t. We are interested in a one-shot approximation of the entire Pareto curve as oppposed to sampling points on it and solving several independent SDPs. Such an approach has been taken before for multi-objective LPs in [84], and for bi-objective polynomial optimization problems in [136].

To get the Pareto curve in one shot, we can solve the following TV-SDP

$$\begin{array}{ll} \max\limits_{x\in\mathbf{L^n}} & \int_0^1 \langle c_1, x(t)\rangle \mathrm{d}t \\ \text{subject to} & \left.\begin{array}{l} \langle c_2, x(t)\rangle \geq t \\ Fx(t) \succeq 0 \end{array}\right\} \quad \forall t \in [0, 1] \text{ a.e..} \end{array} \tag{2.33}$$

If $x \in \mathbf{L^n}$ is any feasible solution to this TV-SDP, then

$$\langle c_1, x(t) \rangle \leq y(t) \; \forall t \in [0,1] \text{ a.e..}$$

In other words, any feasible solution to the TV-SDP in (2.33) gives a lower to the Pareto curve almost every where on $[0,1]$. Furthermore, if x^{opt} is an optimal solution to the same TV-SDP (whose existence is guaranteed by Theorem 3 when $\mathcal{F}$ is compact), then

$$\langle c_1, x^{\text{opt}}(t) \rangle = y(t) \; \forall t \in [0,1] \text{ a.e..}$$

Let $x^d \in \mathbb{R}^n_d[t]$ be an optimal solution to (2.33) when the search space is restricted to polynomials of degree at most d. We know from Theorem 4 that, under the strict feasibility assumption[2] in Definition 2.3.1,

$$\int_0^1 y(t) - \langle c_1, x^d(t) \rangle \, \mathrm{d}t \to 0 \text{ as } d \to \infty.$$

Moreover, the optimal value of the dual problem of the TV-SDP in (2.33) at level d, as described in Section 2.4, gives an upper bound on the area under the Pareto curve. Under the assumption that the set $\mathcal{F}$ is bounded in the infinity norm by γ, then once the constraint $\|x\|_\infty \leq \gamma$ is added to the TV-SDP in (2.33), the optimal values of the associated dual problems converge to the area under the Pareto curve as $d \to \infty$ (see Theorem 7).

As a concrete example of a bi-objective SDP, we consider the Markowitz portfolio selection problem [139]. We model n tradable assets as a nondegenerate n-variate Gaussian random variable with average return $r \in \mathbb{R}^n$ and (positive definite) covariance matrix $\Sigma \in \mathcal{S}^n$. Given the data r and Σ as input, the goal is to choose a portfolio (i.e. an allocation of x_i fraction of our total funds to asset $i \in \{1, \ldots, n\}$) that maximizes the average return $r^T x$ while simultaneously minimizing the variance $x^T \Sigma x$.

We can formulate this problem as a bi-objective optimization problem, with variables

$$\begin{pmatrix} u, & x_1, & \ldots, & x_n \end{pmatrix}^T \in \mathbb{R}^{n+1},$$

constraints

$$x \geq 0, \sum_{i=1}^n x_i \leq 1, \; x^T \Sigma x \leq u,$$

and two objective functions

$$r^T x \text{ and } -u.$$

[2] In this setup, this assumption is equivalent to existence of positive scalar ε and a vector $x^s \in \mathbb{R}^n$ such that $Fx^s \succeq \varepsilon I$ and $\langle c_2, x^s \rangle \geq 1 + \varepsilon$.

The Pareto curve is therefore given by $\{(t, y(t)) \mid t \in [0,1]\}$, where

$$\begin{array}{rll} y(t) := & \max\limits_{x \in \mathbb{R}^n, u \in \mathbb{R}} & r^T x \\ & \text{subject to} & x \geq 0 \\ & & \sum_{i=1}^n x_i \leq 1 \\ & & u \leq t \\ & & \begin{pmatrix} u & x^T \\ x & \Sigma^{-1} \end{pmatrix} \succeq 0. \end{array} \tag{2.34}$$

The TV-SDP in (2.33) that gives this Pareto curve in one shot can therefore be written as

$$\begin{array}{ll} \max\limits_{x \in \mathbf{L^n}, u \in \mathbf{L^1}} & \int_0^1 r^T x(t) \mathrm{d}t \\ \text{subject to} & \\ & \left.\begin{array}{l} x(t) \geq 0 \\ \sum_{i=1}^n x_i(t) \leq 1 \\ u(t) \leq t \\ \begin{pmatrix} u(t) & x(t)^T \\ x(t) & \Sigma^{-1} \end{pmatrix} \succeq 0 \end{array}\right\} \quad \forall t \in [0,1] \text{ a.e.}. \end{array} \tag{2.35}$$

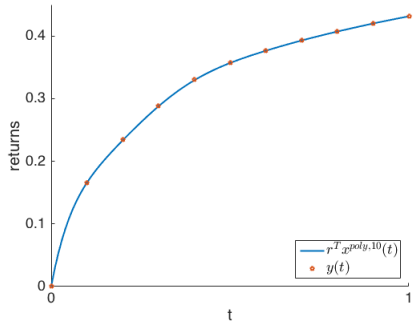
We numerically solve an example with $n = 5$ assets,

$$r = \left(0.4170, 0.7203, 0.0001, 0.3023, 0.1468\right)^T, \Sigma = \begin{pmatrix} 6.0127 & -0.7381 & -0.5441 & -4.9189 & 1.7855 \\ -0.7381 & 9.8904 & -0.7946 & 0.2481 & -5.5214 \\ -0.5441 & -0.7946 & 5.1961 & -3.6240 & 1.5820 \\ -4.9189 & 0.2481 & -3.6240 & 10.4637 & 1.7840 \\ 1.7855 & -5.5214 & 1.5820 & 1.7840 & 15.8475 \end{pmatrix}.$$

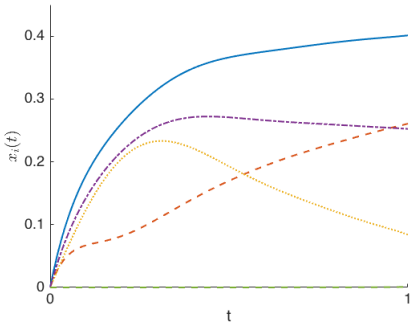
The entries of the vector r were generated independently from the uniform distribution over $[0,1]$. The matrix Σ was obtained by first generating a 5×5 matrix V whose entries were sampled independently from the uniform distribution over $[0,3]$, and then letting $\Sigma = VV^T$.

Using Theorem 6, we solve a semidefinite program that finds the best the best polynomial solution of degree than 10 to the TV-SDP in (2.33). The objective value that we achieve is 0.3210, and the resulting optimal solution $x^{\mathrm{poly},10} \in \mathbb{R}^5_{10}[t]$ is plotted in Figure 2.5b. In Figure 2.5a, we plot $r^T x^{\mathrm{poly},10}(t)$, which is a point-wise lower approximation to the true Pareto curve. We also find eleven equally-spaced points on the exact Pareto curve, by solving the problem in (2.34) at $t \in \{0, 0.1, \ldots, 1\}$. Notice that our approximation to the Pareto curve obtained from the best polynomial solution of degree 10 is almost perfect at these eleven sample points.

To get a formal upper bound on the area enclosed between $\{(t, r^T x^{\mathrm{poly},10}(t)) \mid t \in [0,1]\}$ and the true Pareto curve $\{(t, y(t)) \mid t \in [0,1]\}$, we solve the dual problem (2.15) presented in Section 2.4. After some rewriting, this dual problem at level d is equivalent to the following SDP (cf. Theorem 8):



(a) The return $r^T x^{\text{poly},10}(t)$ obtained by best polynomial solution of degree ≤ 10.



(b) The allocations $x_i^{\text{poly},10}(t)$ for the different assets obtained by the best polynomial solution of degree ≤ 10.

Figure 2.5: The optimal polynomial solution of degree less than 10 and its associated approximation to the Pareto curve for the Markowitz portfolio selection problem.

$$\begin{array}{ll} \max\limits_{x\in\mathbb{R}_d^n[t],u\in\mathbb{R}_d[t]} & \int_0^1 r^T x(t)\mathrm{d}t \\ & \alpha_d^{1*}(x_i) \succeq 0, \qquad \beta_d^{1*}(x_i) \succeq 0 \quad i=1,\ldots,n \\ & \alpha_d^{1*}(1-\sum_{i=1}^n x_i) \succeq 0, \qquad \beta_d^{1*}(1-\sum_{i=1}^n x_i) \succeq 0 \\ \text{subject to} & \alpha_d^{1*}(t-u(t)) \succeq 0, \qquad \beta_d^{1*}(t-u(t)) \succeq 0 \\ & \alpha_d^{n+1*}\begin{pmatrix} u(t) & x(t)^T \\ x(t) & \Sigma^{-1} \end{pmatrix} \succeq 0, \quad \beta_d^{n+1*}\begin{pmatrix} u(t) & x(t)^T \\ x(t) & \Sigma^{-1} \end{pmatrix} \succeq 0. \end{array} \tag{2.36}$$

The optimal value of problem (2.36) with $d = 10$ is equal to 0.3232, which tells us that

$$\int_0^1 \left(y(t) - r^T x^{\text{poly},10}(t)\right)\, \mathrm{d}t \leq \frac{1}{100}\int_0^1 y(t)\, \mathrm{d}t.$$

2.6 Future Research Directions

We end by mentioning a few questions that are left for future research. We believe there is much research to be done to extend some of the fundamental structural results from the continuous linear programming literature (e.g., results related to duality theory or the structure of optimal solutions) to the case of TV-SDPs. As a concrete example, we would be interested in knowing to what extent the duality theory of Pullan [170] can carry over to the TV-SDP setting.

Closer to the focus of this chapter, we have shown in Theorem 4 that under the strict feasibility assumption in Definition 2.3.1, the sequence of objective values of the best polynomial solution of degree d converges to the optimal value of the TV-SDP as $d \to \infty$. If we are interested in a feasible solution with (additive or multiplicative)

error bounded by α, how large should we take d to be as a function of α and other problem parameters? The answer to this question would likely have a dependence on the scalar ε in Definition 2.3.1. Is there an efficient method for obtaining a lower bound on ε, or even checking the strict feasibility assumption? Lastly, we are interested in knowing whether the strict feasibility assumption in Theorem 4 can be weakened, for instance, to existence of a feasible polynomial solution.

Similarly in Theorem 7, we have shown that under a boundedness assumption, the sequence of optimal values of our dual problem at level d converges from above to the optimal value of the TV-SDP. It would be interesting to study the convergence rate of this sequence. We also would like to know if the boundedness assumption is needed for convergence, and whether the bound constraints need to be explicitly added to the TV-SDP as we do now.

Finally, at a more basic level, what is the complexity (in the Turing model of computation) of testing feasibility of a continuous linear program with polynomially-varying data? Here, the maximum degree of the polynomials in the data can either be fixed or part of the input. The reason we do not ask this complexity question for TV-SDPs is that the question is well known to be open even for standard SDPs (see e.g. [66]).

Chapter 3

On Sum of Squares Representation of Convex Forms and Generalized Cauchy-Schwarz Inequalities

3.1 Introduction and Main Result

The set $H_{n,k}$ of homogeneous real polynomials (forms) in n variables and of degree k is a central subject of study in algebraic geometry. When the degree $k =: 2d$ is even, three convex cones inside $H_{n,k}$ have received considerable interest. The cone of *nonnegative* forms

$$P_{n,2d} := \{p \in H_{n,2d} \mid p(\mathbf{x}) \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n\},$$

the cone of *sum of squares (sos)* forms

$$\Sigma_{n,2d} := \{p \in H_{n,2d} \mid p = \sum_i q_i^2 \text{ for some forms } q_i \in H_{n,d}\},$$

and the cone of *convex* forms

$$C_{n,2d} := \{p \in H_{n,2d} \mid \nabla^2 p(\mathbf{x}) \succeq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n\},$$

where $\nabla^2 p(\mathbf{x})$ stands for the Hessian of the form p at $\mathbf{x}$, and the symbol $\succeq$ stands for the partial ordering generated by the cone of positive semidefinite matrices.

The systematic study of the interplay between the cones $P_{n,2d}$ and $\Sigma_{n,2d}$ was undertaken by Hilbert at the end of the nineteenth century, when he showed that these two cones are different unless $n \leq 2$, $2d \leq 2$ or $n = 3, 2d = 4$ [97]. Even though Hilbert's work provided a strategy for constructing nonnegative forms that are not sos for the smallest number of variables and degrees possible (i.e., forms in $P_{3,6} \setminus \Sigma_{3,6}$ and $P_{4,4} \setminus \Sigma_{4,4}$), it took almost eighty years for the first explicit examples of such forms to be found by Motzkin and Robinson [143, 177, 174]. See [175] for a more thorough discussion of the history of this problem.

The relationship between $C_{n,2d}$ and $\Sigma_{n,2d}$ is much more complicated and it was an open problem for some time whether $C_{n,2d} \subseteq \Sigma_{n,2d}$ for all n and d. (The reverse inclusion is of course false; e.g., $x^2y^2 \in \Sigma_{2,4} \setminus C_{2,4}$.) Note however, that we trivially have $C_{n,2d} \subseteq P_{n,2d}$ since a global minimum of a convex form is always at the origin where the form and its gradient vanish.

From an applied and computational perspective, the study of the interplay between the notions of convexity and being a sum of squares could enhance our understanding of existing polynomial optimization algorithms. Concretely, consider the problem of finding the minimum value p^* that a convex (not necessarily homogeneous[1]) polynomial p takes on $\mathbb{R}^n$. To tackle this problem, there exists at least two distinct families of algorithms. The first family is comprised of variants of *descent methods* which are guaranteed to converge to a global minimizer in the presence of convexity, but fail to take advantage of the algebraic structure of the objective function given by the polynomial p. Alternatively, the well-known machinery of "sum of squares relaxation" [158, 157] offers a hierarchy of *semidefinite programs* whose optimal values monotonically converge to p^*. For instance, the first level of this hierarchy gives a lower bound p^{sos} on p^*:

$$p^{\text{sos}} := \max_{\gamma \in \mathbb{R}} \gamma \text{ s.t. } p - \gamma \text{ is sos.}$$

However, this second family of algorithms makes no assumptions about convexity of the polynomial p, and as a result, does not explicitly exploit this property. Now, if we knew a priori that the convex polynomial $p - \gamma$ is sos whenever it is nonnegative, then the first level of this relaxation becomes *exact*; i.e., $p^* = p^{\text{sos}}$.

Blekherman has recently shown that for any fixed degree $2d \geq 4$, as the number of variables n goes to infinity, one encounters considerably more convex forms than sos forms [44]. Remarkably however, there is not a single known example of a convex form that is not sos. Due to Hilbert's characterization of the cases of equality between the cone of nonnegative forms and the cone of sos forms, the smallest cases where one could have hope of finding such an example correspond to quaternary quartics ($n = 4, 2d = 4$) and ternary sextics ($n = 3, 2d = 6$). The goal of this chapter prove that no such example exists among quaternary quartics.

Theorem 3.1.1. *Every convex quarternary quartic is sos, i.e.,* $C_{4,4} \subseteq \Sigma_{4,4}$.

Furthermore, we show that if a conjecture of Blekherman related to the so-called Cayley-Bacharach relations is true, no convex form which is not sos can exist among ternary sextics either, i.e., $C_{3,6} \subseteq \Sigma_{3,6}$.

A possible plan of attack to show that a convex form is sos is to show that it is *sos-convex*. This concept, introduced by Helton and Nie [93], is an algebraic sufficient condition for convexity which also implies the property of being sos. This plan would not be successful for our purposes however, since there exist explicit

[1] While the properties of being nonnegative and sum of squares are preserved under the homogenization operation $p(\mathbf{x}) \to y^{\deg(p)} p(\frac{\mathbf{x}}{y})$, the property of convexity is not in general. For instance, the polynomial $x^2 - 1$ is convex, but its homogenization $x^2 - y^2$ is not.

examples of convex forms that are not sos-convex for both cases $n = 4, 2d = 4$ and $n = 3, 2d = 6$ [9]. In fact, the problem of characterizing for which degrees $2d$ and number of variables n sos-convexity is also a necessary condition for convexity, has been completely solved in [9]. The authors prove that this is the case if and only if $n \leq 2$, $2d \leq 2$ or $n = 3, 2d = 4$, i.e., the same cases for which $P_{n,2d} = \Sigma_{n,2d}$ as characterized by Hilbert, albeit for different reasons.

Our proof strategy relies instead on an equivalence due to Blekherman [45] between the membership $p \in \Sigma_{4,4}$ for a nonnegative form p, and the following bounds on point evaluations of the form p:

$$\sqrt{p(\mathbf{u}_1)} \leq \sum_{i=2}^{8} \sqrt{p(\mathbf{u}_i)} \text{ and } \sqrt{2}\sqrt{|p(\mathbf{z})| + \operatorname{Re}(p(\mathbf{z}))} \leq \sum_{i=3}^{8} \sqrt{p(\mathbf{v}_i)}, \tag{3.1}$$

where the real vectors $\mathbf{v}_i$ and $\mathbf{u}_i$ and the complex vector $\mathbf{z}$ come from intersections of quadratic forms (see theorem 3.4.2 for a more precise statement). This equivalence is explained in section 3.4. We show in section 3.5 that any quaternary quartic form p that satisfies the two inequalities

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^4 \quad Q_p(\mathbf{x}, \mathbf{y}) \leq \sqrt{p(\mathbf{x})p(\mathbf{y})} \quad \text{and} \quad \forall\, \mathbf{z} \in \mathbb{C}^4 \quad |p(\mathbf{z})| \leq Q_p(\mathbf{z}, \bar{\mathbf{z}}), \tag{3.2}$$

where $Q_p(\mathbf{x}, \mathbf{y}) := \frac{1}{12}\mathbf{y}^T \nabla^2 p(\mathbf{x})\mathbf{y}$, also satisfies these bounds. These inequalities can be thought of as a generalization of the Cauchy-Schwarz inequality, valid for any $n \times n$ positive semidefinite matrix Q:

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad \mathbf{x}^T Q \mathbf{y} \leq \sqrt{\mathbf{x}^T Q \mathbf{x} \cdot \mathbf{y}^T Q \mathbf{y}}.$$

We show that convex quaternary quartic forms satisfy the inequalities in eq. (3.2), and are therefore sos. In fact, in section 3.3, we present generalizations of the Cauchy-Schwarz inequality that apply to convex forms of any degree and any number of variables. We believe that these inequalities could be of independent interest. In section 3.6, we discuss a possible extension of our proof technique to the case of ternary sextics.

3.2 Background and Notation

We denote the set of positive natural numbers, real numbers, and complex numbers by $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{C}$ respectively. We denote by $(\mathbf{e_1}, \ldots, \mathbf{e_n})$ the canonical basis of $\mathbb{R}^n$. We denote by i the imaginary number $\sqrt{-1}$, and by $\bar{z}$, $|z|$, $\operatorname{Re}(z)$, and $\operatorname{Im}(z)$ the complex conjugate, the modulus, the real part and the imaginary part of a complex number z respectively.

3.2.1 Notation for differential operators

We denote by $\partial_{\mathbf{u}}$ the partial differentiation operator in the direction of $\mathbf{u} \in \mathbb{C}^n$, i.e., $\partial_{\mathbf{u}} p(\mathbf{x})$ is the limit of the ratio $(p(\mathbf{x}+t\mathbf{u}) - p(\mathbf{x}))/t$ as $t \to 0$ for all n-variate polynomial

functions p and all vectors $\mathbf{x} \in \mathbb{C}^n$. The gradient operator $(\partial_{\mathbf{e_1}}, \ldots, \partial_{\mathbf{e_n}})^T$ is denoted by ∇, the Hessian operator $\nabla\nabla^T$ is denoted by ∇^2, and the Laplacian operator $\partial^2_{\mathbf{e_1}} + \cdots + \partial^2_{\mathbf{e_n}}$ is denoted by Δ. For a form $p \in H_{n,2d}$ and vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{C}^n$, we denote by $p(\partial_{\mathbf{x}_1}, \ldots, \partial_{\mathbf{x}_n})$ the differential operator obtained by replacing the indeterminate x_k with $\partial_{\mathbf{x}_k}$ for $k = 1, \ldots, n$ in the expression $p(x_1, \ldots, x_n)$. We note that taking k partial derivatives of a k-degree form results in a constant function. As a consequence, we consider the quantity $p(\partial_{\mathbf{x}_1}, \ldots, \partial_{\mathbf{x}_n})q$ to be a scalar for all forms p and q in $H_{n,k}$.

3.2.2 Euler's identity

Euler's identity (see e.g., [121]) links the value that a form $p \in H_{n,k}$ takes to its gradient as follows :

$$\forall \mathbf{x} \in \mathbb{R}^n \quad k\, p(\mathbf{x}) = \mathbf{x}^T \nabla p(\mathbf{x}).$$

By applying this identity to the entries of the gradient ∇p, one obtains the following relationship between a form and its Hessian:

$$\forall \mathbf{x} \in \mathbb{R}^n \quad k(k-1)\, p(\mathbf{x}) = \mathbf{x}^T \nabla^2 p(\mathbf{x}) \mathbf{x}.$$

It is readily seen from this identity that every convex form is nonnegative; i.e., for every $d \in \mathbb{N}$, $C_{n,2d} \subseteq P_{n,2d}$.

3.2.3 Tensors and outer product

A tensor of order k is a multilinear form $T : (\mathbb{R}^n)^k \to \mathbb{R}$. The tensor T is called *symmetric* if $T(\mathbf{x}_1, \ldots, \mathbf{x}_k) = T(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k})$ for every $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ and every permutation $(i_1, \ldots, i_k)$ of the set $\{1, \ldots, k\}$. The outer product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\mathbf{x} \otimes \mathbf{y}$. The symmetric outer product $\frac{1}{2}(\mathbf{x} \otimes \mathbf{y} + \mathbf{y} \otimes \mathbf{x})$ of two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\mathbf{x} \cdot \mathbf{y}$. The (symmetric) outer product of a vector $\mathbf{x}$ with itself k times is denoted by $\mathbf{x}^k$. For any tensor T of order k, the quantity $T(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ is a linear function of the outer product $\mathbf{x}_1 \otimes \cdots \otimes \mathbf{x}_k$ of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$. If the tensor T is assumed to be symmetric, then this quantity only depends on the symmetric outer product $\mathbf{x}_1 \ldots \mathbf{x}_k$.

3.2.4 Forms and symmetric tensors

For every form $p \in H_{n,k}$, there exists a unique symmetric tensor T_p of order k such that

$$\forall \mathbf{x} \in \mathbb{R}^n \quad p(\mathbf{x}) = T_p(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{k \text{ times}}).$$

This is known as the *polarization* identity [72]. The tensor T_p is related to the derivatives of the form p via the relation

$$\forall \mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n \quad k!\, T_p(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \partial_{\mathbf{x}_1} \ldots \partial_{\mathbf{x}_k} p, \tag{3.3}$$

and is related to the coefficients of the form p via the identity

$$p_{i_1,\dots,i_n} = \binom{k}{i_1,\dots,i_k} T_p(\underbrace{\mathbf{e_1},\dots,\mathbf{e_1}}_{i_1 \text{ times}},\dots,\underbrace{\mathbf{e_n},\dots,\mathbf{e_n}}_{i_n \text{ times}}), \tag{3.4}$$

where $p_{i_1,\dots,i_n}$ is the coefficient multiplying the monomial $x_1^{i_1}\dots x_n^{i_n}$ in p.

When $k =: 2d$ is even, we define the polynomial Q_p via the formula

$$\forall \mathbf{x},\mathbf{y} \in \mathbb{R}^n \quad Q_p(\mathbf{x},\mathbf{y}) := T_p(\underbrace{\mathbf{x},\dots,\mathbf{x}}_{d \text{ times}},\underbrace{\mathbf{y},\dots,\mathbf{y}}_{d \text{ times}}). \tag{3.5}$$

We call the polynomial Q_p the *biform* associated to p. We note that Q_p is a form of degree $2d$ in the $2n$ variables $(\mathbf{x},\mathbf{y})$ that is homogeneous of degree d in $\mathbf{x}$ (resp. $\mathbf{y}$) when $\mathbf{y}$ (resp. $\mathbf{x}$) is fixed.

3.2.5 Inner product on $H_{n,2d}$

We equip the vector space $H_{n,2d}$ with the following inner product

$$\forall p,q \in H_{n,2d} \quad \langle p,q\rangle := p(\partial_{\mathbf{e_1}},\dots,\partial_{\mathbf{e_n}})q,$$

the so-called *Fischer* inner product [77]. This inner product can also be expressed in a more symmetric way in terms of the coefficients of the forms p and q as follows

$$\forall p,q \in H_{n,2d} \quad \langle p,q\rangle = (2d)! \sum_{i_1+\dots+i_n=2d} \binom{2d}{i_1,\dots,i_n}^{-1} p_{i_1,\dots,i_n}\, q_{i_1,\dots,i_n}.$$

By the Riesz representation theorem, for every linear form $\ell : H_{n,2d} \to \mathbb{R}$, there exists a unique form $p \in H_{n,2d}$ satisfying

$$\forall q \in H_{n,2d} \quad \ell(q) = p(\partial_{\mathbf{e_1}},\dots,\partial_{\mathbf{e_n}})q,$$

and we write $\ell = p(\partial_{\mathbf{e_1}},\dots,\partial_{\mathbf{e_n}})$.

A particularly important special case of linear forms is given by tensor evaluations. The linear form given by $p \mapsto T_p(\mathbf{x}_1,\dots,\mathbf{x}_{2d})$ for some fixed vectors $\mathbf{x}_1,\dots,\mathbf{x}_{2d}$ is identified with the differential operator $1/(2d)!\,\partial_{\mathbf{x}_1}\dots\partial_{\mathbf{x}_{2d}}$. For instance,

- The point evaluation map at $\mathbf{x} \in \mathbb{C}^n$ given by $p \mapsto p(\mathbf{x})$ is equal to the differential operator $\frac{1}{(2d)!}\partial_{\mathbf{x}}^{2d}$.
- The map $p \mapsto Q_p(\mathbf{x},\mathbf{y})$ is equal to the differential operator $\frac{1}{(2d)!}\partial_{\mathbf{x}}^d\partial_{\mathbf{y}}^d$ for all vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{C}^n$.
- The map $p \mapsto Q_p(\mathbf{z},\bar{\mathbf{z}})$ is equal to the differential operator $\frac{1}{(2d)!}(\partial_{\mathbf{x}}^2+\partial_{\mathbf{y}}^2)^d$ for any vector $\mathbf{z}$ in $\mathbb{C}^n$ whose real and imaginary parts are given by $\mathbf{x}$ and $\mathbf{y}$. This follows from the fact that $\partial_{\mathbf{z}} = \partial_{\mathbf{x}} + i\partial_{\mathbf{y}}$ and $\partial_{\bar{\mathbf{z}}} = \partial_{\mathbf{x}} - i\partial_{\mathbf{y}}$.

3.2.6 Convex duality

We denote the dual of a convex cone $\Omega \subseteq H_{n,2d}$ by

$$\Omega^* := \{\ell : H_{n,2d} \to \mathbb{R} \mid \ell \text{ is linear and for all } p \in \Omega \quad \ell(p) \geq 0\}.$$

Recall that $C^*_{n,2d} = \text{cone}\{\ell_{\mathbf{x},\mathbf{y}} \mid \mathbf{x}, \mathbf{y} \in \mathbb{R}^n\}$, where $\ell_{\mathbf{x},\mathbf{y}}(p) := \mathbf{y}^T \nabla^2 p(\mathbf{x}) \mathbf{y}$ and $\text{cone}(S)$ denotes the conic hull of a set S [176]. By using the pairing between linear forms acting on the vector space $H_{n,2d}$ and elements of this vector space described in section 3.2.5, we can write $C^*_{n,2d} = \text{cone}\{(\partial_{\mathbf{y}})^2 (\partial_{\mathbf{x}})^{2d-2} \mid \mathbf{x}, \mathbf{y} \in \mathbb{R}^n\}$. For example, when $n = 2$, if we denote $\partial_{\mathbf{e_1}}$ and $\partial_{\mathbf{e_2}}$ by ∂_x and ∂_y respectively, then

$$C^*_{2,2d} = \left\{ \sum_{k=1}^{N} (\alpha_k \partial_x + \beta_k \partial_y)^2 (\gamma_k \partial_x + \delta_k \partial_y)^{2d-2} \mid N \in \mathbb{N} \text{ and } \alpha_k, \beta_k, \gamma_k, \delta_k \in \mathbb{R} \right\}. \tag{3.6}$$

3.3 Generalized Cauchy-Schwarz Inequalities for Convex Forms

The Cauchy-Schwarz inequality states that for any $n \times n$ positive semidefinite matrix Q,

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad \mathbf{x}^T Q \mathbf{y} \leq \sqrt{\mathbf{x}^T Q \mathbf{x} \cdot \mathbf{y}^T Q \mathbf{y}}.$$

When the vectors $\mathbf{x}$ and $\mathbf{y}$ are complex and conjugate of each other, i.e. when $\mathbf{x} = \bar{\mathbf{y}} =: \mathbf{z}$, the inequality reverses as follows:

$$\forall \mathbf{z} \in \mathbb{C}^n \quad \sqrt{\mathbf{z}^T Q \mathbf{z} \cdot \bar{\mathbf{z}}^T Q \bar{\mathbf{z}}} \leq \mathbf{z}^T Q \bar{\mathbf{z}}.$$

This inequality is well-defined since the quantity appearing on the left-hand side is a nonnegative number as $\mathbf{z}^T Q \mathbf{z} \cdot \bar{\mathbf{z}}^T Q \bar{\mathbf{z}} = |\mathbf{z}^T Q \mathbf{z}|^2$, and the complex number on the right-hand side is a real number because it is equal to its conjugate.

The condition that the matrix Q is positive semidefnite can be restated equivalently in terms of convexity of the quadratic form $p(\mathbf{x}) := \mathbf{x}^T Q \mathbf{x}$. In the following theorem, we present a generalization of these inequalities for convex forms of higher degree.

Theorem 3.3.1 (Generalized Cauchy-Schwarz inequalities (GCS))**.** *For any convex form p in n variables and of degree $2d$, we have*

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad Q_p(\mathbf{x}, \mathbf{y}) \leq A_d \sqrt{p(\mathbf{x}) p(\mathbf{y})}, \tag{3.7}$$

and

$$\forall \mathbf{z} \in \mathbb{C}^n \quad |p(\mathbf{z})| \leq B_d \, Q_p(\mathbf{z}, \bar{\mathbf{z}}), \tag{3.8}$$

where Q_p is the biform associated with p and defined in eq. (3.5), and A_d and B_d are universal positive constants depending only on the degree $2d$.

Proof. See section 3.3.1. □

We emphasize that the constants A_d and B_d appearing in the GCS inequalities depend only on the degree $2d$, and not on the number of variables n. Furthermore, for the purposes of this chapter, we need to find the smallest constants that make these GCS inequalities hold for quartic and sextic forms, i.e., when $d = 2$ and $d = 3$. This motivates the following definitions for all $d \in \mathbb{N}$:

$$\begin{aligned} A_d^* &:= \inf_{A \geq 0} A \quad \text{s.t. } \forall n \in \mathbb{N},\ \forall p \in C_{n,2d},\ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad Q_p(\mathbf{x}, \mathbf{y}) \leq A\sqrt{p(\mathbf{x})p(\mathbf{y})}, \\ B_d^* &:= \inf_{B \geq 0} B \quad \text{s.t. } \forall n \in \mathbb{N},\ \forall p \in C_{n,2d},\ \forall \mathbf{z} \in \mathbb{C}^n \qquad\qquad |p(\mathbf{z})| \leq BQ_p(\mathbf{z}, \bar{\mathbf{z}}). \end{aligned} \tag{3.9}$$

The "inf" in these definitions is actually a "min" since the inequality symbol "$\leq$" appearing in the GCS inequalities is not strict. Moreover, the constants A_d^* and B_d^* are bounded below by 1 for all $d \in \mathbb{N}$. This is easily seen by, e.g., taking $n = 1$, $\mathbf{x} = \mathbf{y} = \mathbf{z} = 1$, and considering the (univariate) convex form $p(x) := x^{2d}$.

Before further discussion of the values of the constants A_d^* and B_d^*, we present in the following two remarks new interpretations of the GCS inequalities that do no involve the biform Q_p.

Remark 4. *In view of the identification of differential operators with linear forms discussed in section 3.2.5, the generalized Cauchy-Schwarz inequality in eq. (3.7) can be written in terms of mixed derivatives as follows:*

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad \partial_{\mathbf{x}}^d \partial_{\mathbf{y}}^d p \leq A_d \sqrt{\partial_{\mathbf{x}}^{2d} p \cdot \partial_{\mathbf{y}}^{2d} p},$$

where p is any convex form of degree $2d$. Similarly, the second Generalized Cauchy-Schwarz inequality in eq. (3.8) can be written as

$$\forall \mathbf{z} \in \mathbb{C}^n \quad |\partial_{\mathbf{z}}^{2d} p| \leq B_d\, \partial_{\mathbf{z}}^d \partial_{\bar{\mathbf{z}}}^d\, p,$$

for any convex form p of degree $2d$. If we denote by $\mathbf{x}$ and $\mathbf{y}$ the real and imaginary parts of the vector $\mathbf{z}$, the same inequality reads

$$|\partial_{\mathbf{z}}^{2d} p| \leq B_d\, (\partial_{\mathbf{x}}^2 + \partial_{\mathbf{y}}^2)^d\, p.$$

Remark 5. *For all forms $p \in H_{n,2d}$, and for all complex vectors $\mathbf{z} \in \mathbb{C}^n$ whose real and imaginary parts are given by $\mathbf{x}$ and $\mathbf{y}$, the quantity $Q_p(\mathbf{z}, \bar{\mathbf{z}})$ is proportional to the average of the form p on the ellipse*

$$\{\alpha\mathbf{x} + \beta\mathbf{y} \mid \alpha, \beta \in \mathbb{R} \text{ and } \alpha^2 + \beta^2 \leq 1\}.$$

More precisely, we show in section 3.7.1 the identity

$$Q_p(\mathbf{z}, \bar{\mathbf{z}}) = \frac{4^d(d+1)}{\pi} \binom{2d}{d}^{-1} \iint_{\alpha^2+\beta^2 \leq 1} p(\alpha\mathbf{x} + \beta\mathbf{y})\, \mathrm{d}\alpha\mathrm{d}\beta. \tag{3.10}$$

The generalized Cauchy-Schwarz inequality in (3.8) can thus be equivalently written as

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad |p(\mathbf{x} + i\mathbf{y})| \leq B'_d \iint_{\alpha^2+\beta^2 \leq 1} p(\alpha\mathbf{x} + \beta\mathbf{y}) \, \mathrm{d}\alpha \mathrm{d}\beta,$$

for any convex form $p \in H_{n,2d}$, where $B'_d := \frac{4^d(d+1)}{\binom{2d}{d}\pi} B_d$.

theorem 3.3.1 is equivalent to the statement that the constants A^*_d and B^*_d are finite for all positive integers d. The following theorem strengthens this claim.

Theorem 3.3.2 (Optimal constants in the GCS inequalities)**.** *For all positive integers d,*

$$B^*_d = \frac{\binom{2(d-1)}{d-1}}{d}.$$

Moreover, $A^*_1 = A^*_2 = A^*_3 = 1$, A^*_4 is an algebraic number of degree 3, and for all even integers $d \geq 4$, $A^*_d > 1$. More generally, for every positive integer d, the constant A^*_d is the optimal value of an (explicit) semidefinite program.

Proof. See section 3.3.2. □

Remark 6. *The quantity $\frac{\binom{2(d-1)}{d-1}}{d}$ is known as the d^{th} Catalan number [99].*

3.3.1 Proof of the generalized Cauchy-Schwarz inequalities

In this section, we will show that the GCS inequalities are, at heart, linear inequalities about bivariate convex forms. This observation will eventually lead to a simple proof of theorem 3.3.1.

The next lemma leverages the homogeneity properties of the elements of $H_{n,2d}$ to linearize inequalities eqs. (3.7) and (3.8).

Lemma 3.3.3. *For all $n, d \in \mathbb{N}$, for any positive constants A_d and B_d, and for any nonnegative form $p \in P_{n,2d}$,*
(i) the form p satisfies the inequality in eq. (3.7) with constant A_d if and only if

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad 2Q_p(\mathbf{x}, \mathbf{y}) \leq A_d \left(p(\mathbf{x}) + p(\mathbf{y})\right), \tag{3.11}$$

(ii) the form $p \in P_{n,2d}$ satisfies the inequality in eq. (3.8) with constant B_d if and only if

$$\forall \mathbf{z} \in \mathbb{C}^n \quad \mathrm{Re}\left(p(\mathbf{z})\right) \leq B_d Q_p(\mathbf{z}, \bar{\mathbf{z}}). \tag{3.12}$$

Proof. Fix positive integers n and d, positive scalars A_d and B_d, and let $p \in P_{n,2d}$. Let us prove part (i) of the lemma first, i.e., that the form p satisfies eq. (3.7) if and only if it satisfies eq. (3.11). The "only if" direction can be easily seen from the inequality

$$\forall a, b \geq 0 \quad \sqrt{ab} \leq \frac{a+b}{2}.$$

We now turn our attention to the "if" direction. Applying inequality eq. (3.11) to vectors $\mathbf{x}$ and $\lambda^{\frac{1}{d}}\mathbf{y}$, where λ is a nonnegative scalar, results in

$$2Q_p(\mathbf{x}, \lambda^{\frac{1}{d}}\mathbf{y}) \leq A_d \left(p(\mathbf{x}) + p(\lambda^{\frac{1}{d}}\mathbf{y})\right).$$

By homogeneity, we get that $2\lambda Q_p(\mathbf{x}, \mathbf{y}) \leq A_d \ (p(\mathbf{x}) + \lambda^2 p(\mathbf{y}))$. In other words, for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the univariate polynomial $f(\lambda) := A_d \ p(\mathbf{x}) + \lambda^2 \ A_d \ p(\mathbf{y}) - 2\lambda \ Q_p(\mathbf{x}, \mathbf{y})$ is nonnegative on $[0, \infty)$. Since the form p is assumed to be nonnegative, the scalars $p(\mathbf{x})$ and $p(\mathbf{y})$ are nonnegative. If one of these two scalars is zero, or if the scalar $Q_p(\mathbf{x}, \mathbf{y})$ is negative, the inequality in eq. (3.11) follows immediately. Otherwise, the polynomial f has two (complex) roots, whose sum and product are both positive. The polynomial f is therefore nonnegative on $[0, \infty)$ if and only if its roots are equal or are not real. In the first case, the discriminant $Q_p(\mathbf{x}, \mathbf{y})^2 - A_d^2 p(\mathbf{x}) p(\mathbf{y})$ is zero, and in the second case, the discriminant is negative. In both cases, the inequality in eq. (3.11) follows.

We now prove part (ii) of the lemma, i.e., that the form p satisfies eq. (3.8) if and only if it satisfies eq. (3.12). Again, it is straightforward to see why the "only if" part is true, so we only prove the "if" part. Let us assume that p satisfies inequality eq. (3.12) with constant B_d, and let $\mathbf{z}$ be an arbitrary complex vector in $\mathbb{C}^n$. Let $\mathbf{z}' = e^{i\theta}\mathbf{z}$, where $\theta := \frac{\arg(p(\mathbf{z}))}{2d}$ is chosen so that $p(\mathbf{z}')$ is a nonnegative scalar. By homogeneity, we have

$$|p(\mathbf{z})| = \mathrm{Re}\left(p(\mathbf{z}')\right) \text{ and } Q_p(\mathbf{z}, \bar{\mathbf{z}}) = Q_p(\mathbf{z}', \bar{\mathbf{z}}').$$

Applying inequality eq. (3.12) to $\mathbf{z}'$ leads to $|p(\mathbf{z})| \leq B_d \ Q_p(\mathbf{z}, \bar{\mathbf{z}})$, which is the desired result. □

We now show that it suffices to prove the GCS inequalities for convex forms in 2 variables. For this purpose, notice that for any n-variate form p of degree $2d$ and for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the restriction of p to the plane spanned by the vectors $\mathbf{x}$ and $\mathbf{y}$

$$q(x, y) := p(x\mathbf{x} + y\mathbf{y}). \tag{3.13}$$

retains all relevant information for the inequality eq. (3.7). Indeed,

$$p(\mathbf{x}) = q(\mathbf{e_1}), p(\mathbf{y}) = q(\mathbf{e_2}) \text{ and } Q_p(\mathbf{x}, \mathbf{y}) = Q_q(\mathbf{e_1}, \mathbf{e_2}),$$

where $\mathbf{e_1}^T = (1, 0)$ and $\mathbf{e_2}^T = (0, 1)$. Moreover, for any complex vector $\mathbf{z} = \mathbf{x} + i\mathbf{y}$, we have $p(\mathbf{z}) = q(\mathbf{e_1} + i\mathbf{e_2}), Q_p(\mathbf{z}, \bar{\mathbf{z}}) = Q_q(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2})$, and thus all the quantities appearing in the inequality eq. (3.8) only depend on p through its two-dimensional restriction q as well.

The form q defined in eq. (3.13) is bivariate and of the same degree as p. Furthermore, the form q is convex if p is. The proof of theorem theorem 3.3.1 therefore reduces to showing existence of two constants A_d and B_d indexed by $d \in \mathbb{N}$, such that

all bivariate convex forms q of degree $2d$ satisfy the inequalities

$$2Q_q(\mathbf{e_1}, \mathbf{e_2}) \leq \ A_d \ (q(\mathbf{e_1}) + q(\mathbf{e_2}))\,, \tag{3.14}$$

and

$$\mathrm{Re}\,(q(\mathbf{e_1} + i\mathbf{e_2})) \leq B_d\, Q_q(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2}). \tag{3.15}$$

We now show that these inequalities follow from the following simple lemma, whose proof is delayed until the end of the section.

Lemma 3.3.4. *Let Ω be a closed cone in $H_{n,2d}$, and let ℓ be a linear form defined on Ω that satisfies*

$$\forall p \in \Omega \ [\ell(p) = 0 \implies p = 0] \qquad \text{and} \qquad \ell \geq 0 \text{ on } \Omega. \tag{3.16}$$

Then the set $\Omega_\ell := \{p \in \Omega \mid \ell(p) \leq 1\}$ is compact.

Proof of theorem 3.3.1. Fix $d \in \mathbb{N}$, and define two linear forms ℓ and s acting on $q \in C_{2,2d}$ as $\ell(q) := q(\mathbf{e_1}) + q(\mathbf{e_2})$ and $s(q) := Q_q(\mathbf{z}, \bar{\mathbf{z}})$, where $\mathbf{z} = \mathbf{e_1} + i\mathbf{e_2}$. We start by showing that the linear forms ℓ and s satisfy the condition in eq. (3.16) with $\Omega = C_{2,2d}$. If $q \in C_{2,2d}$, then q is nonnegative and therefore $\ell(q) \geq 0$. Moreover, because of the relationship between $Q_q(\mathbf{z}, \bar{\mathbf{z}})$ and the integral of q described in eq. (3.10), it is clear that $s(q) \geq 0$ as well. Now assume that a form $q \in C_{2,2d}$ satisfies $\ell(q) = 0$. Since q is nonnegative, we have $q(\mathbf{e_1}) = q(\mathbf{e_2}) = 0$. By convexity, the restriction of the function q to the segment linking $\mathbf{e_1}$ to $\mathbf{e_2}$ is identically zero. By homogeneity, q must be identically zero. Similarly, if $s(q) = 0$, then by eq. (3.10), the average of q on the unit disk is zero, and since the form q is assumed to be nonnegative, it must be identically 0.

By lemma 3.3.4, the following two sets must therefore be compact:

$$L := \{q \in C_{2,2d} \mid q(\mathbf{e_1}) + q(\mathbf{e_2}) \leq 1\}, \quad S := \{q \in C_{2,2d} \mid Q_p(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2}) \leq 1\}.$$

Let $\|\cdot\|$ be any norm on the vector space $H_{2,2d}$ and define the scalars α, β as follows:

$$\alpha := \sup_{q \in L} \|q\| \text{ and } \beta := \sup_{q \in H_{2,2d}} \frac{2Q_q(\mathbf{e_1}, \mathbf{e_2})}{\|q\|}.$$

We will show that inequality eq. (3.14) holds with constant $A_d = \alpha\beta$. Note that α is finite because L is compact, and β is finite because the map $q \mapsto Q_q(\mathbf{e_1}, \mathbf{e_2})$ is a linear function over a finite dimensional space. Let $q \in C_{2,2d}$ and assume that q is not zero, so that the scalar $\ell(q)$ is positive. On the one hand, we have $2Q_q(\mathbf{e_1}, \mathbf{e_2}) \leq \beta\|q\|$. On the other hand, $\|q\| \leq \alpha\ell(q)$ because the form $\frac{q}{\ell(q)}$ is in the set L. We have just shown that $2Q_q(\mathbf{e_1}, \mathbf{e_2}) \leq \alpha\beta\ell(q)$, which concludes the proof of inequality (3.14). A similar argument shows the existence of a finite constant B_d for which (3.15) hold. □

Proof of lemma 3.3.4. Let Ω and ℓ be as in the statement of the lemma, and let $\|\cdot\|$ be any norm of $H_{n,2d}$. It is clear that the set Ω_ℓ is a closed set as it is the intersection of a half space with Ω.

Suppose it is not bounded, i.e., suppose there exists a sequence $(q^{(k)})_k$ of Ω_ℓ such that $\|q^{(k)}\| \to \infty$ as $k \to \infty$. By taking a subsequence if necessary, we can assume that the sequence $(q^{(k)})_k$ does not contain zero. The sequence $\left(\frac{q^{(k)}}{\|q^{(k)}\|}\right)_k$ lives in the cone Ω, and is bounded (by one), so it admits a converging subsequence. Let $q_\infty \in \Omega$ denote its limit. Since the function ℓ is bounded by 1 on Ω_ℓ, the ratio $\frac{\ell(q^{(k)})}{\|q^{(k)}\|}$ tends to 0 as $k \to \infty$, and therefore $\ell(q_\infty) = 0$. By eq. (3.16), the form q_∞ is itself identically zero. Yet, $\|q_\infty\| = 1$, which is a contradiction. □

3.3.2 Values of the optimal constants A_d^* and B_d^* defined in eq. (3.9)

Fix $d \in \mathbb{N}$. We have shown in the previous section that

$$\begin{aligned} A_d^* &= \min A \text{ s.t. } \quad \forall q \in C_{2,2d} \quad 2Q_q(\mathbf{e_1}, \mathbf{e_2}) \leq A\left(q(\mathbf{e_1}) + q(\mathbf{e_2})\right), \\ B_d^* &= \min B \text{ s.t. } \quad \forall q \in C_{2,2d} \quad \operatorname{Re}\left(q(\mathbf{e_1} + i\mathbf{e_2})\right) \leq BQ_q(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2}). \end{aligned} \tag{3.17}$$

This formulation is useful for finding lower bounds on the constants A_d^* and B_d^*. Indeed, to show that $A_d^* > A$ for some scalar A , it suffices to exhibit a convex bivariate form q that satisfies $2Q_q(\mathbf{e_1}, \mathbf{e_2}) > A\left(q(\mathbf{e_1}) + q(\mathbf{e_2})\right)$. A similar statement can be made for B_d^* as well.

To find upper bounds on the constants A_d^* and B_d^*, we take a dual approach. For all scalars A and B, we define the linear forms

$$\ell_A := A(\partial_x^{2d} + \partial_y^{2d}) - 2\partial_x^d \partial_y^d, \tag{3.18}$$

and

$$s_B := B(\partial_x^2 + \partial_y^2)^d - \operatorname{Re}\left((\partial_x + i\partial_y)^{2d}\right). \tag{3.19}$$

Because of our discussion in section 3.2.5, the constants A_d^* and B_d^* can be found by solving the following optimization problems, dual to the optimization problems in eq. (3.17).

$$\begin{aligned} A_d^* &= \min A \quad \text{s.t.} \quad \ell_A \in C_{2,2d}^*, \\ B_d^* &= \min B \quad \text{s.t.} \quad s_B \in C_{2,2d}^*. \end{aligned} \tag{3.20}$$

In other words, in order to prove that $A_d^* \leq A$ for some scalar A, one has to show that ℓ_A can be decomposed as in eq. (3.6). An identical statement can be made for B_d^* here too.

Values of the optimal constants A_d^* defined in eq. (3.9)

We will show in this section that the optimization problems in eq. (3.20) are tractable. The following theorem shows that convex bivariate forms are also sos-convex.

Theorem 3.3.5. *[9, Theorem 5.1] A bivariate form $q(x, y) = \sum_{i=0}^{2d} q_i x^i y^{2d-i}$ is convex if and only if it is sos-convex, i.e., if there exists a positive semidefinite $2d \times 2d$ matrix*

d	1	2	3	4	5	6	7	8
A_d^*	1.000	1.000	1.000	1.011	1.000	1.061	1.000	1.048

Table 3.1: Approximation of the value of the constant A_d^* defined in eq. (3.9) obtained by numerically solving the SDP in eq. (3.20)

Q such that

$$\forall x, y \in \mathbb{R}, \forall \mathbf{u} \in \mathbb{R}^2 \quad \mathbf{u}^T \nabla^2 q(x,y) \mathbf{u} = z^T Q z, \tag{3.21}$$

where $z^T := (u_1 x^{d-1}, u_1 x^{d-2} y, \ldots, u_1 y^{d-1}, u_2 x^{d-1}, u_2 x^{d-2} y, \ldots, u_2 y^{d-1})$ is the vector of monomials in the variables x, y, u_1, u_2.

By expanding both sides of eq. (3.21) and matching the coefficients of the polynomials that appear on both sides, we obtain an equivalent system of linear equations involving the coefficients of the form q and the entries of the matrix Q. What we have just shown is that

$$C_{2,2d} = \{q \in H_{2,2d} \mid \exists Q \succeq 0 \text{ s.t. } q \text{ and } Q \text{ satify the linear equations in eq. (3.21)}\}.$$

This set is a *projected spectrahedron*, i.e., it is defined via linear equations and linear matrix inequalities. The class of projected spectrahedra is stable by taking the convex dual, so $C_{2,2d}^*$ is also a projected spectrahedron. Optimizing linear functions over such sets (or their duals) is therefore a semidefinite program (SDP). For any fixed integer d, the optimization problems in eq. (3.20) characterizing the constant A_d^* and B_d^* are therefore SDPs. Semidefinite programming is a well-studied subclass of convex optimization problems that can be solved to arbitrary accuracy in polynomial time. We report in table 3.1 the values of A_d^* ($d = 1, \ldots, 8$) to 4 digits of accuracy obtained using the solver MOSEK [22].

Note that in practice, numerical software will only return an *approximation* of the optimal solution to an SDP. Such approximations can nevertheless be useful as they help formulate a "guess" for what the exact solution might be, especially if the solution sought contains only rational numbers (with small denominators). In particular, the identities below, which are trivial to verify, were obtained by rounding solutions obtained from a numerical SDP solver.

$$\begin{aligned}
\partial_x^2 + \partial_y^2 - 2\partial_x\partial_y &= (\partial_x - \partial_y)^2, \\
\partial_x^4 + \partial_y^4 - 2\partial_x^2\partial_y^2 &= (\partial_x - \partial_y)^2(\partial_x + \partial_y)^2, \\
\partial_x^6 + \partial_y^6 - 2\partial_x^3\partial_y^3 &= \frac{1}{2}(\partial_x - \partial_y)^2(\partial_x^4 + \partial_y^4 + (\partial_x + \partial_y)^4).
\end{aligned}$$

The right-hand side (and therefore, the left-hand side) of each one of these identities is in $C_{2,2d}^*$ for $d = 1, 2, 3$ respectively by eq. (3.6). These identities constitute a formal proof that $A_1^*, A_2^*, A_3^* \leq 1$. The reverse inequality $A_d^* \geq 1$ valid for all integers d was

already proved in section 3.3. We therefore have

$$A_1^*, A_2^*, A_3^* = 1$$

We also note that few algebraic methods have been developed for solving SDPs exactly, especially for problems of small sizes [15, 96]. For instance, we were able to solve the SDP in (3.20) characterizing A_d^* for $d = 4$. The key steps in this computation are (i) exploiting the symmetries of the problem to reduce the size of the SDP [81], (ii) formulating the corresponding *Karush-Kuhn-Tucker* (KKT) equations (see, e.g., [20]) and (iii) solving these polynomial equations using variable elimination techniques. [2] The value of A_4^* is given by

$$\frac{1}{70}\,\omega^{\frac{1}{3}} + \frac{128}{15}\,\omega^{-\frac{1}{3}} + \frac{11}{35}, \text{ where } \omega := 14336 + i\frac{14336\sqrt{3}}{9}.$$

Unlike the constants A_d^* for $d \leq 3$, the constant A_4^* is not equal to 1. In fact A_4^* is not even a rational number, but an algebraic number of degree three with minimal polynomial given by

$$t^3 - \frac{33}{35}\,t^2 - \frac{17}{245}\,t + \frac{13}{42875}.$$

In section 3.7.2, we prove that $A_d^* > 1$ whenever d is an even integer larger than 4.

Conjecture and open problem. Supported by the numerical evidence in table 3.1, we conjecture that the constant A_d^* defined in (3.9) is equal to 1 when the integer d is odd, and we leave open the problem of finding the exact value of A_d^* for even integers d larger than 4.

Exact values of the optimal constants B_d^* defined in eq. (3.9)

The goal of this section is to show that

$$\forall d \in \mathbb{N} \quad B_d^* = \frac{\binom{2(d-1)}{d-1}}{d}.$$

The following proposition shows that for all positive integers d, $s_B \in C_{2,2d}^*$ for $B = \frac{\binom{2(d-1)}{d-1}}{d}$, where s_B is defined in eq. (3.19), and therefore, $B_d^* \leq \frac{\binom{2(d-1)}{d-1}}{d}$.

Proposition 3.3.6. *For all positive integers d, for all $x, y \in \mathbb{R}$,*

$$\frac{\binom{2(d-1)}{d-1}}{d}(x^2+y^2)^d - \mathrm{Re}\left((x+iy)^{2d}\right) = \frac{4^d}{2d}\sum_{k=0}^{d-1}\left(-s_k x + c_k y\right)^2 \left(c_k x + s_k y\right)^{2d-2}, \quad (3.22)$$

where $c_k = \cos(\frac{k\pi}{2d})$ and $s_k = \sin(\frac{k\pi}{2d})$ for $k = 0, 1, \ldots, 2d-1$.

Proof. Identity eq. (3.22) is homogeneous in $\mathbf{x}^T = (x, y)$. It is therefore sufficient to prove that it holds when $\mathbf{x}$ is a unit vector. Let $\mathbf{x}$ be such a vector, and let us write

[2] The curious reader is referred to this Sage notebook describing these steps in more details [75].

$x = \cos(\theta)$ and $y = \sin(\theta)$ for some $\theta \in \mathbb{R}$. Then identity eq. (3.22) becomes

$$\forall \theta \in \mathbb{R} \quad \frac{\binom{2(d-1)}{d-1}}{d} - \cos(2d\theta) = \frac{4^d}{2d} \sum_{k=0}^{d-1} \sin^2\left(\frac{k\pi}{2d} - \theta\right) \cos^{2d-2}\left(\frac{k\pi}{2d} - \theta\right).$$

We give the proof of this trigonometric identity below. To simplify notations, let

$$f_d(\theta) := \sum_{j=0}^{d-1} \sin^2\left(\frac{j\pi}{d} - \theta\right) \cos^{2d-2}\left(\frac{j\pi}{d} - \theta\right).$$

For $j \in \mathbb{N}$, let $r_j := e^{-i\frac{j\pi}{d}}$. Using the fact that

$$\cos\left(\frac{j\pi}{d} - \theta\right) = \frac{e^{i\theta} r_j + e^{-i\theta} \bar{r_j}}{2} \text{ and } \sin\left(\frac{j\pi}{d} - \theta\right) = \frac{e^{i\theta} r_j - e^{-i\theta} \bar{r_j}}{2i},$$

we get that

$$f_d(\theta) = -\frac{1}{2^{2d}} \sum_{j=0}^{d-1} \left(e^{i\theta} r_j - e^{-i\theta} \bar{r_j}\right)^2 \left(e^{i\theta} r_j + e^{-i\theta} \bar{r_j}\right)^{2d-2}.$$

By expanding and exchanging the order of the summation, we get

$$f_d(\theta) = -\frac{1}{2^{2d}} \sum_{h=0}^{2d-2} \binom{2d-2}{h} \left((e^{i\theta})^{2h} \sum_{j=0}^{d-1} r_j^{2h} + (e^{i\theta})^{2h-4} \sum_{j=0}^{d-1} r_j^{2h-4} - 2(e^{i\theta})^{2h-2} \sum_{j=0}^{d-1} r_j^{2h-2}\right).$$

We now use the following simple fact about the sum of the k^{th} powers of roots of unity

$$\forall k \in \mathbb{N} \quad \sum_{j=0}^{d-1} r_j^{2k} = \begin{cases} d & \text{if } d \text{ divides } k \\ 0 & \text{otherwise} \end{cases}$$

to get[3]

$$f_d(\theta) = -\frac{d}{2^{2d}} \sum_{h=0}^{2d-2} \binom{2d-2}{h} \left(e^{2i\theta h} 1_{\{d \mid h\}} + e^{2i(h-2)\theta} 1_{\{d \mid h-2\}} - 2e^{2i(h-1)\theta} 1_{\{d \mid h-1\}}\right),$$

and therefore

$$f_d(\theta) = \frac{2d}{2^{2d}} \left(\frac{\binom{2d-2}{d-1}}{d} - \cos(2d\theta)\right).$$

□

[3]The notation $1_{d|h}$ stands for 1 if d divides h and 0 otherwise.

Let us now show that for all $d \in \mathbb{N}$, the constant B_d^* is bounded below by $\frac{\binom{2d-2}{d-1}}{d}$. To do so, we exhibit a family of nonzero bivariate convex forms $(q_d)_{d\in\mathbb{N}}$ that satisfy

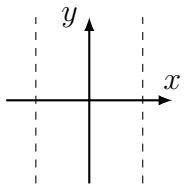
$$\forall d \in \mathbb{N} \quad \mathrm{Re}\left(q_d(\mathbf{e_1} + i\mathbf{e_2})\right) = \frac{\binom{2d-2}{d-1}}{d}\, Q_{q_d}(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2}).$$

We plot in fig. 3.1 the 1-level sets of the polynomials q_d for $d = 1, \ldots, 4$.

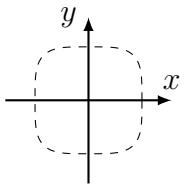
Proposition 3.3.7. *For every positive integer d, the form q_d defined by*

$$q_d(x, y) := \mathrm{Re}\left((x + i\, y)^{2d}\right) + (2d - 1)(x^2 + y^2)^d \tag{3.23}$$

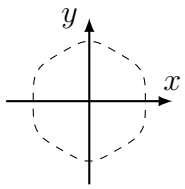
is convex and satisfies $\mathrm{Re}\left(q_d(\mathbf{e_1} + i\mathbf{e_2})\right) = \frac{\binom{2d-2}{d-1}}{d}\, Q_{q_d}(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2})$.



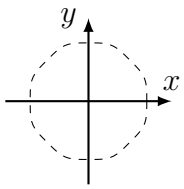
$q_1(x, y) = 2x^2$



$q_2(x, y) = 4x^4 + 4y^4$



$q_3(x, y) = 6x^6 + 30x^2y^4 + 4y^6$



$q_4(x, y) = 8x^8 + 112x^4y^4 + 8y^8$

Figure 3.1: Plot of the 1-level sets of the forms q_d defined in (3.23) for $d = 1, \ldots, 4$. These forms saturate the generalized Cauchy-Schwarz inequality in (3.8).

To give the proof of proposition 3.3.7, it will be convenient for us to switch to *polar coordinates* $(r, \theta) \in [0, \infty) \times [0, 2\pi)$ defined by $x = r\cos(\theta)$ and $y = r\sin(\theta)$. More explicitly, for every $k \in \mathbb{N}$, every bivariate form $p \in H_{2,k}$ can be expressed in polar coordinates as $p(x, y) = r^k f(\theta)$, where f is a polynomial expression in $\cos(\theta)$ and $\sin(\theta)$. In particular, the function f is differentiable infinitely many times. The following lemma gives the expressions of the Hessian and Laplacian operators in polar coordinates.

Lemma 3.3.8 (Hessian and Laplacian in polar coordinates)**.** *The Hessian and Laplacian of a form $p \in H_{2,k}$, whose expression in polar coordinates is $p(x, y) = r^k f(\theta)$, are given by*

$$\nabla^2 p(x, y) = r^{k-2}\ \left(k(k-1)f(\theta)\mathbf{e_{rr}} + (k-1)f'(\theta)\mathbf{e_{r\theta}} + (k + f''(\theta))\,\mathbf{e}_{\theta\theta}\right),$$

$$\Delta p(x, y) = r^{k-2}\ \left(k^2 f(\theta) + f''(\theta)\right),$$

where $\mathbf{e_r} := \begin{pmatrix}\cos(\theta)\\ \sin(\theta)\end{pmatrix}, \mathbf{e}_\theta := \begin{pmatrix}-\sin(\theta)\\ \cos(\theta)\end{pmatrix}$, *and* $\mathbf{e_{rr}} = \mathbf{e_r}\mathbf{e_r}^T, \mathbf{e_{r\theta}} = \mathbf{e_r}\mathbf{e_\theta}^T + \mathbf{e_\theta}\mathbf{e_r}^T, \mathbf{e_{\theta\theta}} = \mathbf{e_\theta}\mathbf{e_\theta}^T$.

Proof. Let p and f be as in the statement of the lemma. Recall that the gradient operator ∇ can be written in polar coordinates as follows

$$\nabla = \frac{\partial}{\partial r}\mathbf{e_r} + \frac{1}{r}\frac{\partial}{\partial\theta}\mathbf{e}_\theta.$$

The Hessian operator $\nabla^2 = \nabla \cdot \nabla^T$ is thus given by

$$\nabla^2 = \frac{\partial^2}{\partial r^2}\mathbf{e_{rr}} + \frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial\theta}\right)\mathbf{e_{r\theta}} + \left(\frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}\right)\mathbf{e_{\theta\theta}}.$$

Note that taking the derivative of a form of degree $k' \geq 1$ with respect to r is equivalent to multiplying by $\frac{k'}{r}$. The Hessian operator, when applied to the k-degree form p, can thus be simplified further to

$$\nabla^2 p(x,y) = r^{k-2}\ \left(k(k-1)f(\theta)\mathbf{e_{rr}} + (k-1)f'(\theta)\mathbf{e_{r\theta}} + (k + f''(\theta))\,\mathbf{e_{\theta\theta}}\right).$$

The Laplacian Δp is given by the trace of the matrix $\nabla^2 p$. Since the trace of both matrices $\mathbf{e_{rr}}$ and $\mathbf{e_{\theta\theta}}$ is one and the trace of $\mathbf{e_{r\theta}}$ is zero, we get that

$$\Delta p(x,y) = r^{k-2}\ \left(k^2 f(\theta) + f''(\theta)\right).$$

□

Proof of proposition 3.3.7. Fix a positive integer d and let us prove that the form q_d defined in eq. (3.23) is convex. Note that we can express q_d in polar coordinates as follows,

$$q_d(x,y) = \mathrm{Re}\,(r^{2d}e^{i2d\theta} + (2d-1)r^{2d}).$$

Using lemma 3.3.8, we get that

$$\nabla^2 r^{2d} = r^{2d-2}\,(2d(2d-1)\mathbf{e_{rr}} + 2d\mathbf{e_{\theta\theta}})$$

and

$$\nabla^2\left(r^{2d}e^{i2d\theta}\right) = 2d(2d-1)r^{2d-2}e^{i2d\theta}\,(\mathbf{e_{rr}} + i\mathbf{e_{r\theta}} - \mathbf{e_{\theta\theta}}).$$

By summing the previous two equations term by term and taking the real part, we get

$$\nabla^2 q_d(x,y) = 2d(2d-1)r^{2d-2}\begin{pmatrix}\cos(2d\theta) + 2d - 1 & -\sin(2d\theta)\\ -\sin(2d\theta) & -\cos(2d\theta) + 1\end{pmatrix}.$$

The trace of the matrix in the right-hand side of this equation is $(2d)^2(2d-1)r^{2d-2}$, and its determinant is given by $(2d(2d-1)r^{2d-2})^2(2d-2)(1+\cos(2d\theta))$. Both the trace and the determinant of the Hessian matrix of q_d are thus clearly nonnegative. This proves that this Hessian matrix is positive semidefinite and that the form q_d is convex.

Let us now compute $\mathrm{Re}\,(q_d(\mathbf{z}))$ and $Q_{q_d}(\mathbf{z},\bar{\mathbf{z}})$ where $\mathbf{z} = \mathbf{e_1} + i\mathbf{e_2}$. By plugging $x = 1$ and $y = i$ in the right-hand side of the identity

$$\mathrm{Re}\,((x+iy)^{2d}) = \sum_{k=0}^{d} \binom{2d}{2k} x^{2d-2k}(iy)^{2k},$$

we get that $q_d(\mathbf{z}) = \sum_{k=0}^{d} \binom{2d}{2k} = 2^{2d-1}$.

We now compute $Q_{q_d}(\mathbf{z},\bar{\mathbf{z}})$. Because of the identification between linear forms and differential operators introduced in section 3.2.5, this task is equivalent to computing $\Delta^d q_d$. On the one hand, the function $f(x,y) := (x+iy)^{2d}$ is holomorphic when viewed as a function of the complex variable $z = x + iy$, therefore

$$\Delta^d\,\mathrm{Re}\,((x+iy)^{2d}) = 0.$$

On the other hand, by using lemma 3.3.8 again, we get for every positive integer k, $\Delta r^{2k} = 4k^2 r^{2k-2}$, and by immediate induction, $\Delta^d r^{2d} = 2^{2d} d!^2$. Overall, we get $\Delta^d q_d = (2d-1)2^{2d} d!^2$, and therefore

$$Q_{q_d}(\mathbf{z},\bar{\mathbf{z}}) = 2^{2d-1}\frac{d}{\binom{2d-2}{d-1}}.$$

In conclusion, we have just proved that $\mathrm{Re}\,(q_d(\mathbf{z})) = \frac{\binom{2d-2}{d-1}}{d} Q_{q_d}(\mathbf{z},\bar{\mathbf{z}})$. □

3.4 What Separates the Sum of Squares Cone from the Nonnegative Cone

In [45], the author offers a complete description of the hyperplanes separating sos forms from non-sos forms inside the cone of nonnegative quaternary quartics. We include the high level details of that description here to make this article relatively self-contained.

If a form $h \in P_{4,4}$ is not sos, then there exists a subset $V = \{\mathbf{v}_1,\ldots,\mathbf{v}_8\}$ of $\mathbb{C}^4$ and complex numbers $a_1,\ldots,a_8 \in \mathbb{C}\setminus\{0\}$ that certify that fact in the sense that[4]

$$\sum_{i=1}^{8} a_i p(\mathbf{v}_i) \geq 0 \quad \forall p \in \Sigma_{4,4}, \tag{3.24}$$

but $\sum_{i=1}^{8} a_i h(\mathbf{v}_i) < 0$ [45, Theorem 1.2]. Let us now explain where the set V and the scalar a_i come from. The points in V are the common zeros to three linearly independent quadratic forms $q_i(\mathbf{x}) = \mathbf{x}^T Q_i \mathbf{x}$, where the Q_i are 4×4 real symmetric

[4]The left-hand side of eq. (3.24) is real because the vectors v_i and the scalars a_i come in complex conjugate pairs (see the discussion that follows (3.26).)

matrices and $i = 1, 2, 3$ [45, Lemma 2.9]. Equivalently,

$$V = \{\mathbf{x} \in \mathbb{C}^4 \mid q_1(\mathbf{x}) = q_2(\mathbf{x}) = q_3(\mathbf{x}) = 0\} = \{\mathbf{v}_1, \ldots, \mathbf{v}_8\}.$$

Define $V^2 := \{\mathbf{v}\mathbf{v}^T \mid \mathbf{v} \in V\}$. The eight elements of V^2 live in the 10-dimensional vector space of symmetric 4×4 matrices, and they are all orthogonal to the three-dimensional vector space spanned by Q_1, Q_2, and Q_3. A simple dimension counting argument tells us that there must exist a linear relationship between the vectors $\mathbf{v}_i$ of the form

$$\sum_{i=1}^{8} \mu_i \mathbf{v}_i \mathbf{v}_i^T = 0, \tag{3.25}$$

for some $\mu_1, \ldots, \mu_8 \in \mathbb{C}$. In fact, this relationship between the $\mathbf{v}_i$ is unique (up to scaling). Furthermore, all the scalars μ_i must be nonzero. This is known as the Cayley-Bacharach relation [73]. We assume from now on that all the μ_i have norm 1 (after possibly scaling the vectors $\mathbf{v}_i$.)

Now that we have characterized the evaluation points $\mathbf{v}_i$, let us turn our attention to the scalars a_i in eq. (3.24). These scalars should satisfy [45, Theorem 6.1 and Theorem 7.1]

$$\sum_{i=1}^{8} \frac{1}{a_i} = 0. \tag{3.26}$$

We now need to distinguish between the case where all the elements of V are real (i.e., $V \subset \mathbb{R}^4$) and the case where they are not. In the former case, all the scalars μ_i must be real, exactly one of the scalars a_i must be negative and the rest should be positive [45, Theorem 6.1]. By reordering the indices if necessary, we assume $a_1 < 0$ and $a_i > 0$ for $i > 1$. By scaling all the scalars a_i, we assume

$$\frac{1}{a_1} = -\sum_{i=2}^{8} \frac{1}{a_i} = -1, \tag{3.27}$$

in which case inequality eq. (3.24) reads

$$p(\mathbf{v}_1) \leq \sum_{i=2}^{8} a_i p(\mathbf{v}_i). \tag{3.28}$$

In the case where one of the vectors $\mathbf{v}_i$ is not real, it is proven in [45, Corollary 4.4] that V could be taken so that exactly two of the vectors $\mathbf{v}_i$ are not real, in which case they (and their coefficients μ_i) should be conjugate of each other. Again, up to reordering, we can assume that $\mathbf{v}_1 := \mathbf{z}$ is not real, $\mathbf{v}_2 = \bar{\mathbf{z}}$, $\mu_1 = \bar{\mu}_2$ and the rest of the vectors $\mathbf{v}_i$ and scalars μ_i are real. By scaling, we can further assume that

$$\frac{1}{a_1} + \frac{1}{\bar{a}_1} = -\sum_{i=3}^{8} \frac{1}{a_i} = -1. \tag{3.29}$$

In this case, the inequality in eq. (3.24) reads

$$a_1 p(\mathbf{z}) + \bar{a}_1 p(\bar{\mathbf{z}}) + \sum_{i=3}^{8} a_i p(\mathbf{v}_i) \geq 0. \tag{3.30}$$

We present the following simple lemma (whose proof can be found in section 3.7.3) that will let us rewrite the inequality in eq. (3.24) without refering to the scalars a_i.

Lemma 3.4.1. *For all nonnegative scalars $x_2, \ldots, x_n$, the maximum of the quantity $\sum_{i=2}^{n} a_i x_i$ over all positive scalars $a_2, \ldots, a_n$ satisfying $\sum_{i=2}^{n} \frac{1}{a_i} = 1$ is $(\sum_{i=2}^{n} \sqrt{x_i})^2$. Furthermore, for any complex number z, the maximum value of the quantity $az + \bar{a}\bar{z}$ over all complex numbers a satisfying $\frac{1}{a} + \frac{1}{\bar{a}} = 1$ is $2(|z| + \mathrm{Re}\,(z))$.*

Indeed, lemma 3.4.1 shows that a form p satisfies inequality eq. (3.28) for every $a_1, \ldots, a_8 \in \mathbb{R}$ satisfying eq. (3.27) if and only if

$$p(\mathbf{v}_1) \leq \left(\sum_{i=2}^{8} \sqrt{p(\mathbf{v}_i)} \right)^2,$$

and the same form satisfies Inequality eq. (3.30) for every $a_1 \in \mathbb{C}$ and $a_3 \ldots, a_8 \in \mathbb{R}$ satisfying eq. (3.29) if and only if

$$2(|p(\mathbf{z})| + \mathrm{Re}\,(p(\mathbf{z}))) \leq \left(\sum_{i=3}^{8} \sqrt{p(\mathbf{v}_i)} \right)^2.$$

We summarize the result of this section in the following theorem.

Theorem 3.4.2 ([45])**.** *A nonnegative quaternary quartic form p is sos if and only if both of the following conditions hold.*

- *For every $\mathbf{v}_1, \ldots, \mathbf{v}_8 \in \mathbb{R}^4$ and $\alpha_2, \ldots, \alpha_8 \in \{-1, 1\}$ such that $\mathbf{v}_1 \mathbf{v}_1^T = \sum_{i=2}^{8} \alpha_i \mathbf{v}_i \mathbf{v}_i^T$,*

$$p(\mathbf{v}_1) \leq \left(\sum_{i=2}^{8} \sqrt{p(\mathbf{v}_i)} \right)^2. \tag{3.31}$$

- *For every $\mathbf{z} \in \mathbb{C}^4$, for every $\mathbf{v}_3, \ldots, \mathbf{v}_8 \in \mathbb{R}^4$, and for every $\alpha_3, \ldots, \alpha_8 \in \{-1, 1\}$ such that $\mathbf{z}\mathbf{z}^T + \bar{\mathbf{z}}\bar{\mathbf{z}}^T = \sum_{i=3}^{8} \alpha_i \mathbf{v}_i \mathbf{v}_i^T$,*

$$2(|p(\mathbf{z})| + \mathrm{Re}\,(p(\mathbf{z}))) \leq \left(\sum_{i=3}^{8} \sqrt{p(\mathbf{v}_i)} \right)^2. \tag{3.32}$$

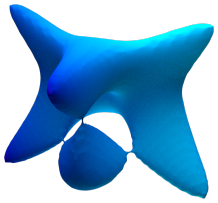


Figure 3.2: The set $\{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid p(x_1, x_2, x_3, 1) = 1\}$, i.e., the section of the 1-level set of the polynomial p (defined in eq. (3.33)) with the hyperplane $\{x_4 = 1\}$.

Example 1. *Let us use theorem 3.4.2 to prove that the following quaternary quartic form*

$$p(\mathbf{x}) = \sum_{i=1}^{4} x_i^4 + \sum_{\substack{1 \le i,j,k \le 4 \\ i \ne j, i \ne k, j \ne k}} x_i^2 x_j x_k + 4 x_1 x_2 x_3 x_4, \tag{3.33}$$

whose 1-level set is depicted in fig. 3.2, is not sos. Take V to be the set of 8 elements given by

$$V := \{-1, 1\} \times \{-1, 1\} \times \{-1, 1\} \times \{1\},$$

and partition it as $V = V^+ \cup V^-$, where V^+ (resp. V^-) is the subset of elements V whose entries sum to an even (resp. odd) number. Up to scaling, the unique linear relationship satisfied by the elements of V is given by

$$\sum_{\mathbf{v} \in V^+} \mathbf{v}\mathbf{v}^T - \sum_{\mathbf{v} \in V^-} \mathbf{v}\mathbf{v}^T = 0.$$

Let $\mathbf{v}_1 \in V$ stand for the vector $(1, 1, 1, 1)^T$, and denote the rest of the elements of V by $\mathbf{v}_2, \ldots, \mathbf{v}_8$. It is easy to check that

$$p(\mathbf{v}_1) = 32 \text{ and } p(\mathbf{v}_i) = 0 \text{ for } i = 2, \ldots, 8.$$

Therefore, p does not satisfy requirement eq. (3.31) in theorem 3.4.2, and is not sos as a result.

3.5 Proof of the Main Theorem

In this section we prove that $C_{4,4} \subseteq \Sigma_{4,4}$. Our plan of action is to show that any quaternary quartic form p that satisfies the following generalized inequality:

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \quad Q_p(\mathbf{x}, \mathbf{y}) \le \sqrt{p(\mathbf{x}) p(\mathbf{y})} \tag{3.34}$$

and

$$\forall \mathbf{z} \in \mathbb{C}^n \quad |p(\mathbf{z})| \le Q_p(\mathbf{z}, \bar{\mathbf{z}}), \tag{3.35}$$

must satisfy the requirements eq. (3.31) and eq. (3.32) that appear in theorem 3.4.2, and hence must be sos. The containment $C_{4,4} \subseteq \Sigma_{4,4}$ then follows since convex qua-

ternary quartics satisfy the generalized Cauchy-Schwarz inequalities with constants $A_2^* = B_2^* = 1$ by theorem 3.3.2.

Let p be a quaternary quartic form satisfying the inequalities in eqs. (3.34) and (3.35), and let us prove that p satisfies both requirements appearing in theorem 3.4.2.

The first requirement in (3.31). Let $\mathbf{v}_1, \dots, \mathbf{v}_8 \in \mathbb{R}^4$ and $\alpha_2, \dots, \alpha_8 \in \{-1, 1\}$ such that $\mathbf{v}_1\mathbf{v}_1^T = \sum_{i=2}^8 \alpha_i \mathbf{v}_i\mathbf{v}_i^T$. Using the tensor notation developed in section 3.2.3, this is equivalent to $\mathbf{v}_1^2 = \sum_{i=2}^8 \alpha_i \mathbf{v}_i^2$. Squaring[5] both sides of this equation leads to

$$\mathbf{v}_1^4 = \sum_{i=2}^{8} \alpha_i\alpha_j \mathbf{v}_i^2\mathbf{v}_j^2.$$

Recall that the biform $(\mathbf{x}, \mathbf{y}) \mapsto Q_p(\mathbf{x}, \mathbf{y})$ defined in eq. (3.5) can be seen as a linear function of the symmetric outer product $\mathbf{x}^2\mathbf{y}^2$. We conclude that

$$p(\mathbf{v}_1) = \sum_{2\leq i,j\leq 8} \alpha_i\alpha_j Q_p(\mathbf{v}_i, \mathbf{v}_j).$$

Using eq. (3.34), we know that $|Q_p(\mathbf{v}_i, \mathbf{v}_j)| \leq \sqrt{p(\mathbf{v}_i)p(\mathbf{v}_j)}$, and therefore

$$p(\mathbf{v}_1) \leq \sum_{2\leq i,j\leq 8} \sqrt{p(\mathbf{v}_i)p(\mathbf{v}_j)}.$$

The second requirement in (3.32). Let $\mathbf{v}_3, \dots, \mathbf{v}_8 \in \mathbb{R}^4$, $\alpha_3, \dots, \alpha_8 \in \{1, -1\}$ and $\mathbf{z} \in \mathbb{C}^4$ such that $\mathbf{z}\mathbf{z}^T + \bar{\mathbf{z}}\bar{\mathbf{z}}^T = \sum_{i=3}^8 \alpha_i \mathbf{v}_i\mathbf{v}_i^T$. Squaring both side of the equation and applying the biform Q_p as before gives:

$$\sum_{3\leq i,j\leq 8} \alpha_i\alpha_j Q_p(\mathbf{v}_i, \mathbf{v}_j) = p(\mathbf{z}) + p(\bar{\mathbf{z}}) + 2Q_p(\mathbf{z}, \bar{\mathbf{z}}) = 2\,\mathrm{Re}\,(p(\mathbf{z})) + 2Q_p(\mathbf{z}, \bar{\mathbf{z}}).$$

On the one hand, using eq. (3.35), we know that $|p(\mathbf{z})| \leq Q_p(\mathbf{z}, \bar{\mathbf{z}})$, so

$$2(\mathrm{Re}\,(p(\mathbf{z})) + |p(\mathbf{z})|) \leq 2\,\mathrm{Re}\,(p(\mathbf{z})) + 2Q_p(\mathbf{z}, \bar{\mathbf{z}}).$$

On the other hand, by eq. (3.34),

$$\sum_{3\leq i,j\leq 8} \alpha_i\alpha_j Q_p(\mathbf{v}_i, \mathbf{v}_j) \leq \sum_{3\leq i,j\leq 8} \sqrt{p(\mathbf{v}_i)p(\mathbf{v}_j)}.$$

In conclusion, $2(|p(\mathbf{z})| + \mathrm{Re}\,(p(\mathbf{z}))) \leq \sum_{i,j=3}^8 \sqrt{p(\mathbf{v}_i)p(\mathbf{v}_j)}$.

[5] The square of a vector $\mathbf{v}$ is simply the outer product of the vector with itself.

3.6 Remarks on the Case of Ternary Sextics

It is natural to ask whether our proof can be extended to show that convex ternary sextics are also sos. theorem 3.4.2 for instance generalizes in a straightforward fashion.

Theorem 3.6.1 ([45]). *A nonnegative ternary sextic form p is sos if and only if both of the following conditions hold.*

- *For every $\mathbf{v}_1, \dots, \mathbf{v}_9 \in \mathbb{R}^3$ and $\alpha_2, \dots, \alpha_9 \in \{-1, 1\}$ such that $\mathbf{v}_1^3 = \sum_{i=2}^9 \alpha_i \mathbf{v}_i^3$,*

$$p(\mathbf{v}_1) \leq \left(\sum_{i=2}^{9} \sqrt{p(\mathbf{v}_i)} \right)^2. \tag{3.36}$$

- *For every $\mathbf{z} \in \mathbb{C}^3$, for every $\mathbf{v}_3, \dots, \mathbf{v}_9 \in \mathbb{R}^3$, and for every $\alpha_3, \dots, \alpha_9 \in \{-1, 1\}$ such that $\mathbf{z}^3 + \bar{\mathbf{z}}^3 = \sum_{i=3}^9 \alpha_i \mathbf{v}_i^3$,*

$$2(|p(\mathbf{z})| + \operatorname{Re}(p(\mathbf{z}))) \leq \left(\sum_{i=3}^{9} \sqrt{p(\mathbf{v}_i)} \right)^2. \tag{3.37}$$

In order for us to follow the same proof strategy that applies to quaternary quartics to the set of ternary sextics, we would take an arbitrary convex ternary sextic and try to show that it satisfies both requirements appearing in the previous Theorem. The first requirement is easily dealt with since sextics satisfy the generalized Cauchy-Schwarz inequality appearing in eq. (3.7) with a constant A_3^* equal to 1 (similar to the quartics case). Sextics on the other hand satisfy eq. (3.8) only with a constant B_3^* strictly larger than 1 (as opposed to $B_2^* = 1$ for quartics). This proves to be the main obstacle preventing us from showing that convex ternary sextics satisfy the second requirement in theorem 3.6.1. In [45, Conjecture 7.3], the author conjectures that this second requirement is actually not needed, in which case our proof strategy would succeed.

3.7 Omitted proofs

3.7.1 Proof of identity (3.10)

Let q be a $2d$-degree form in n variables and let $\mathbf{x}, \mathbf{y}$ be two vectors in $\mathbb{R}^n$. By considering the restriction $(x, y) \mapsto q(x\mathbf{x} + y\mathbf{y})$ of the form q to the plane spanned by $\mathbf{x}$ and $\mathbf{y}$ if necessary, we can assume without loss of generality that $n = 2$, $\mathbf{x} = \mathbf{e_1}$, and $\mathbf{y} = \mathbf{e_2}$. As a consequence, it suffices to prove that the identity

$$Q_q(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2}) = \frac{4^d(d+1)}{\pi} \binom{2d}{d}^{-1} \iint_{x^2+y^2 \leq 1} q(x, y)\, \mathrm{d}x\mathrm{d}y$$

holds for all bivariate convex forms q of degree $2d$. This identity will follow from the following lemma.

Lemma 3.7.1. *For $k \in \mathbb{N}$, any form p in $H_{2,k}$ satisfies*

$$\iint_{x^2+y^2\leq 1} \Delta p(x,y)\ \mathrm{d}x\mathrm{d}y = k(k+2) \iint_{x^2+y^2\leq 1} p(x,y)\ \mathrm{d}x\mathrm{d}y.$$

Indeed, using this lemma inductively on the iterates $q, \Delta q, \dots, \Delta^{d-1}q$, we get

$$\iint_{x^2+y^2\leq 1} \Delta^d q(x,y)\ \mathrm{d}x\mathrm{d}y = 4^d(d+1)d!^2 \iint_{x^2+y^2\leq 1} q(x,y)\ \mathrm{d}x\mathrm{d}y.$$

Since $\Delta^d q$ is a constant and the area of the unit disk is π, we get that

$$\Delta^d q = \frac{4^d(d+1)d!^2}{\pi} \iint_{x^2+y^2\leq 1} q(x,y)\ \mathrm{d}x\mathrm{d}y.$$

Recall from section 3.2.5 that $Q_q(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2}) = \frac{1}{(2d)!}\Delta^d q$, and therefore

$$Q_q(\mathbf{e_1} + i\mathbf{e_2}, \mathbf{e_1} - i\mathbf{e_2}) = \frac{4^d(d+1)}{\pi}\binom{2d}{d}^{-1} \iint_{x^2+y^2\leq 1} q(x,y)\ \mathrm{d}x\mathrm{d}y,$$

which concludes the proof.

Proof of lemma 3.7.1. Fix $k \in \mathbb{N}$ and a form $p \in H_{2,k}$. Denote by $\mathcal{D}$ (resp. $\partial\mathcal{D}$) the unit disk (resp. unit circle). The well-known divergence theorem states that

$$\iint_{\mathcal{D}} \Delta p(x,y)\ \mathrm{d}x\mathrm{d}y = \oint_{\partial\mathcal{D}} \begin{pmatrix} x \\ y \end{pmatrix}^T \nabla p(x,y),$$

where $\oint_{\partial\mathcal{D}}$ stands for the line integral over $\partial\mathcal{D}$. Euler's identity shows that the integrand on the right-hand side of the previous equation is $kp(x,y)$, and therefore

$$\iint_{\mathcal{D}} \Delta p(x,y)\ \mathrm{d}x\mathrm{d}y = k \oint_{\partial\mathcal{D}} p(x,y).$$

Exploiting the fact that the function p is homogeneous of degree k again to relate the integral on $\mathcal{D}$ to the line integral over $\partial\mathcal{D}$ (see [31, Corollary 1]) leads to

$$\oint_{\partial\mathcal{D}} p(x,y) = (k+2) \iint_{\mathcal{D}} p(x,y)\ \mathrm{d}x\mathrm{d}y,$$

which concludes the proof. □

3.7.2 Proof that the constant A_d^* defined in (3.9) is larger than 1 for all even integers $d \geq 4$

In this section, we will show that for all even integers $d \geq 4$, there exists a convex bivariate form p_d of degree $2d$ that satisfies $p_d(1,0) = p_d(0,1) = 1$ and $Q_{p_d}(\mathbf{e_1}, \mathbf{e_2}) > 1$. This shows that $A_d^* > 1$.

Fix an integer $d \geq 4$ and let $p_d := s + \alpha_d\, q$, where

$$s(x,y) := \frac{(x+y)^{2d} + (x-y)^{2d}}{2}, q(x,y) := \sum_{k=1}^{d-1} x^{2k} y^{2d-2k},$$

and α_d is a positive constant defined explicitly in eq. (3.38). Note that $p_d(1,0) = p_d(1,0) = 1$ and $Q_{p_d}(\mathbf{e_1}, \mathbf{e_2}) = 1 + \frac{\alpha_d}{\binom{2d}{d}} > 1$.

It remains to prove that the form p_d is convex. The idea of the proof is as follows. On the one hand, the Hessian of the form s is positive definite everywhere except on the two lines $y = \pm x$, where it is only positive semidefinite. On the other hand, the Hessian of the form q is positive definite on the two lines $y = \pm x$. By picking α_d to be small enough, we can therefore make the form p_d convex.

More formally, by homogeneity, it suffices to prove that the Hessian of p is positive semidefinite on the circle $\mathcal{S} := \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 = 2\}$. Let us now examine the Hessians of the forms s and q individually. The Hessian of s is given by

$$\nabla^2 s(x,y) = d(2d-1) \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} (x+y)^{2d-2} & 0 \\ 0 & (x-y)^{2d-2} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The matrix $\nabla^2 s(x,y)$ is positive definite for every $(x,y) \in \mathcal{S}$ except on the four points $X := \{(\pm 1, \pm 1)\}$ where it is only positive semidefinite. We will now prove that the Hessian of q is positive definite on X. A simple computation shows that

$$\nabla^2 q(1,1) = \nabla^2 q(-1,-1) = \frac{d(d-1)}{3} \begin{pmatrix} 4d-5 & 2d+2 \\ 2d+2 & 4d-5 \end{pmatrix},$$

$$\nabla^2 q(1,-1) = \nabla^2 q(-1,1) = \frac{d(d-1)}{3} \begin{pmatrix} 4d-5 & -2d-2 \\ -2d-2 & 4d-5 \end{pmatrix}.$$

By examining the trace and the determinant of these matrices (which are univariate polynomials in the variable d), we see that they are positive definite if and only if $d \geq \frac{7}{2}$. Let us now partition the circle $\mathcal{S}$ as

$$\mathcal{S} = U \cup (S \setminus U),$$

where U is any open subset of $\mathcal{S}$ containing X on which the matrix $\nabla^2 q$ is positive definite. If we take

$$\alpha_d := \min_{\|\mathbf{u}\|=1, (x,y) \in S \setminus U} \frac{\mathbf{u}^T \nabla^2 s(x,y) \mathbf{u}}{|\mathbf{u}^T \nabla^2 q(x,y) \mathbf{u}|} > 0, \tag{3.38}$$

then the Hessian of the form $p_d := s + \alpha_d q$ is positive semidefinite on $\mathcal{S}$, and the form p_d itself is therefore convex.

3.7.3 Proof of lemma 3.4.1

Let us first prove that for all nonnegative scalars $x_2, \ldots, x_n$, the optimal value of the minimization problem below is equal to $(\sum_{i=2}^n \sqrt{x_i})^2$.

$$\min \sum_{i=2}^n a_i x_i \quad \text{s.t.} \quad a_i > 0 \text{ for } i = 2, \ldots, n \quad \text{and} \quad \sum_{i=2}^n \frac{1}{a_i} = 1.$$

Let γ stand for the optimal value of this optimization problem. Taking $a_i = \left(\sum_{j=2}^n \sqrt{x_j}\right) x_i^{-\frac{1}{2}}$ for $i = 2, \ldots, n$ (with the convention that $0^{-1} = +\infty$) shows that $\gamma \leq (\sum_{i=2}^n \sqrt{x_i})^2$. We now show that $\gamma \geq (\sum_{i=2}^n \sqrt{x_i})^2$. Consider positive scalars $a_2, \ldots, a_n$ satisfying $\sum_{i=2}^n \frac{1}{a_i} = 1$. Note that

$$\sum_{i=2}^n \frac{1}{a_i} = \mathbf{1}^T A^{-1} \mathbf{1},$$

where $\mathbf{1}^T := (1, \ldots, 1) \in \mathbb{R}^{n-1}$ and A is the diagonal $(n-1)\times(n-1)$ matrix with the a_i as diagonal elements. By taking the Schur complement, the inequality $1 - \mathbf{1}^T A^{-1} \mathbf{1} \geq 0$ implies that $A \succeq \mathbf{1}\mathbf{1}^T$. Therefore, by multiplying each side of this matrix inequality by $\mathbf{u}^T := (\sqrt{x_2}, \ldots, \sqrt{x_n})$, we get $(\mathbf{1}^T \mathbf{u})^2 \leq \mathbf{u}^T A \mathbf{u}$, i.e., $(\sum_{i=2}^n \sqrt{x_i})^2 \leq (\sum_{i=2}^n a_i x_i)^2$. In conclusion, $\gamma \geq (\sum_{i=2}^n \sqrt{x_i})^2$.

Let us now prove that for any complex number z,

$$\max_{a \in \mathbb{C}, \frac{1}{a} + \frac{1}{\bar{a}} = 1} az + \bar{a}\bar{z} = 2(|z| + \mathrm{Re}\,(z)).$$

First notice that $a \in \mathbb{C}$ satisfies $\frac{1}{a} + \frac{1}{\bar{a}} = 1$ if and only if a has the form $2\cos(\theta)e^{i\theta}$ for some $\theta \in \mathbb{R}$.

Write $z = |z|e^{i\alpha}$ for some $\alpha \in \mathbb{R}$. Then,

$$\begin{aligned}
\max_\theta \cos(\theta)\,\mathrm{Re}\,(e^{i\theta} z) &= \max_\theta |z| \cos(\theta)\cos(\theta + \alpha) \\
&= \frac{1}{2}|z| \max(\cos(\alpha) + \cos(2\theta + \alpha)) \\
&= \frac{1}{2}|z|(1 + \cos(\alpha)) \\
&= \frac{|z| + \mathrm{Re}\,(z)}{2}.
\end{aligned}$$

The result follows as $az + \bar{a}\bar{z} = 4\,\mathrm{Re}\,(\cos(\theta)e^{i\theta} z)$.

Part II

Semidefinite Programming for Analyzing and Learning Dynamical Systems

Chapter 4

On Algebraic Proofs of Stability for Homogeneous Vector Fields

4.1 Introduction and Outline of Contributions

We are concerned in this chapter with a continuous time dynamical system

$$\dot{x} = f(x), \tag{4.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable and has an equilibrium at the origin, i.e., $f(0) = 0$. The problem of deciding asymptotic stability of equilibrium points of such systems is a fundamental problem in control theory. The goal of this chapter is prove that if f is a homogeneous vector field (see the definition below), then asymptotic stability is equivalent to existence of a Lyapunov function that is the ratio of two polynomials (i.e., a rational function). We also address the computational question of finding such a Lyapunov function in the case where the vector field f is polynomial.

A scalar valued function $p : \mathbb{R}^n \to \mathbb{R}$ is said to be *homogeneous* of degree $d > 0$ if it satisfies $p(\lambda x) = \lambda^d p(x)$ for all $x \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$. Similarly, we say that a vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ is homogeneous of degree $d > 0$ if $f(\lambda x) = \lambda^d f(x)$ for all $x \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$. Homogeneous vector fields have been extensively studied in the literature on nonlinear control; see e.g. [181], [19], [88, Sect. 57], [87], [30],[71], [38], [103], [57], [23], [56], [179], [142], [76], [212], [110]. These systems are not only of interest as is: they can also be used to study properties of related *non-homogeneous* systems. For example, if one can show that the vector field corresponding to the lowest-degree nonzero homogeneous component of the Taylor expansion of a smooth nonlinear vector field is asymptotically stable, then the vector field itself will be locally asymptotically stable.

We recall that the origin of (4.1) is said to be *stable in the sense of Lyapunov* if for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ such that

$$\|x(0)\| < \delta \;\Rightarrow \|x(t)\| < \epsilon, \quad \forall t \geq 0.$$

We say that the origin is *locally asymptotically stable* if it is stable in the sense of Lyapunov and if there exists a scalar $\hat{\delta} > 0$ such that

$$\|x(0)\| < \hat{\delta} \;\Rightarrow\; \lim_{t\to\infty} x(t) = 0.$$

The origin is *globally asymptotically stable* if it is stable in the sense of Lyapunov and $\lim_{t\to\infty} x(t) = 0$ for any initial condition in $\mathbb{R}^n$. A basic fact about homogeneous vector fields is that for these systems the notions of local and global asymptotic stability are equivalent. Indeed, the values that a homogeneous vector field f takes on the unit sphere determines its value everywhere.

It is also well known that the origin of (4.1) is globally asymptotically stable if there exists a continuously differentiable Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}$ which is radially unbounded (i.e., satisfies $V(x) \to \infty$ as $\|x\| \to \infty$), vanishes at the origin, and is such that

$$V(x) > 0 \quad \forall x \neq 0 \tag{4.2}$$

$$-\langle \nabla V(x), f(x)\rangle > 0 \quad \forall x \neq 0. \tag{4.3}$$

Throughout this chapter, whenever we refer to a *Lyapunov function*, we mean a function satisfying the aforementioned properties. We say that V is *positive definite* if it satisfies (4.2). When V is a homogeneous function, the inequality (4.2) can be replaced by

$$V(x) > 0 \quad \forall x \in \mathcal{S}^{n-1},$$

where $\mathcal{S}^{n-1}$ here denotes the unit sphere of $\mathbb{R}^n$. It is straightforward to check that a positive definite homogeneous function is automatically radially unbounded.

The first contribution of this chapter is to show that an asymptotically stable homogeneous and continuously differentiable vector field always admits a Lyapunov function which is a *rational function* (Theorem 4.3.1). This is done by utilizing a well-known result on existence of homogeneous Lyapunov functions [179], [88], [212], [110] and proving a statement on simultaneous approximation of homogeneous functions and their derivatives by homogeneous rational functions (Lemma 4.2.1).

4.1.1 Polynomial vectors fields

We pay special attention in this chapter to the case where the vector field f in (4.1) is polynomial. Polynomial differential equations appear ubiquitously in applications—either as true models of physical systems or as approximations of other families of nonlinear dynamics—and have received a lot of attention in recent years because of the advent of promising analysis techniques using sum of squares optimization [156], [165], [95], [148], [53], [94], [107]. In a nutshell, these techniques allow for an automated search over (a subset of) polynomial Lyapunov functions of bounded degree using semidefinite programming. However, there are comparatively few converse results in

the literature (e.g. those in [160], [151], [11], [155]) on guaranteed existence of such Lyapunov functions.

In [12], the authors prove that there are globally asymptotically stable polynomial vector fields (of degree as low as 2) which do not admit polynomial Lyapunov functions. We show in this chapter that the same example in [12] does not even admit a rational Lyapunov function (Section 4.3.1). This counterexample justifies the homogeneity assumption of our Theorem 4.3.1.

Homogeneous polynomial vector fields of degree 1 are nothing but linear systems. In this case, it is well known that asymptotic stability is equivalent to existence of a (homogeneous) quadratic Lyapunov function (see e.g. [112, Thm. 4.6]) and can be checked in polynomial time. Moving up in the degrees, one can show that homogeneous vector fields of even degree can never be asymptotically stable [88, Sect. 17]. When the degree of f is odd and ≥ 3, testing asymptotic stability of (4.1) is not a trivial problem. In fact, already when the degree of f is equal to 3 (and even if we restrict f to be a gradient vector field), the problem of testing asymptotic stability is known to be strongly NP-hard [3]. This result rules out the possibility of a polynomial time or even pseudo-polynomial time algorithm for this task unless P=NP. One difficulty that arises here is that tests of stability based on linearization fail. Indeed, the linearization of f around the origin gives the identically zero vector field. This means (see e.g. [112, Thm. 4.15]) that homogeneous polynomial vector fields of degree ≥ 3 are never exponentially stable. This fact is independently proven by Hahn in [88, Sect. 17].

Our main contribution in this chapter is to show that a proof of asymptotic stability for a homogeneous polynomial vector field can always be found by *semidefinite programming* (Theorem 4.4.3). This statement follows from existence of a rational Lyapunov function whose numerator is a *strictly sum of squares* homogeneous polynomial (see Section 4.4 for a definition) and whose denominator is an even power of the 2-norm of the state. Our result generalizes the classical converse Lyapunov theorem for linear systems which corresponds to the case where the power of the strictly sum of squares polynomial in the numerator (resp. denominator) is two (resp. zero).

Our next contribution is a negative result: We show in Proposition 4.5.1 that unlike the case of linear systems, for homogeneous polynomial vector fields of higher degree, one cannot bound the degree of the numerator of a rational Lyapunov function as a function of only the degree (or even the degree and the dimension) of the input vector field. We leave open the possibility that the degree of this numerator can be bounded as a *computable* function of the coefficients of the input vector field. Such a statement (if true), together with the fact that semidefinite feasibility problems can be solved in finite time [163], would imply that the question of testing asymptotic stability for homogeneous polynomial vector fields is decidable. Decidability of asymptotic stability for polynomial vector fields is an outstanding open question of Arnlod; see [25], [64], [24].

In Section 4.6, we show a curious advantage that rational Lyapunov functions can sometimes have over polynomial ones. In Proposition 4.6.1, we give a family of homogeneous polynomial vector fields of degree 5 that all admit a low-degree rational Lyapunov function but require polynomial Lyapunov functions of arbitrarily

high degree. We end the chapter with some concluding remarks and future research directions in Section 6.6.

4.2 Approximation of Homogeneous Functions by Rational Functions

For a positive even integer k, let $\mathcal{H}_k(\mathbb{R}^n)$ denote the set of continuously differentiable homogeneous functions $V : \mathbb{R}^n \to \mathbb{R}$ of degree k. For a function $V \in \mathcal{H}_k(\mathbb{R}^n)$, we define the norm $\|.\|_{\mathcal{H}}$ as

$$\|V\|_{\mathcal{H}} = \max\left\{\max_{x\in\mathcal{S}^{n-1}} |V(x)|, \max_{x\in\mathcal{S}^{n-1}} \|\nabla V(x)\|_2\right\}.$$

We prove in this section that homogeneous rational functions are dense in $\mathcal{H}_k(\mathbb{R}^n)$ for the norm $\|.\|_{\mathcal{H}}$. We remark that there is an elegant construction by Peet [160] that approximates derivatives of any function that has continuous mixed derivatives of order n by derivatives of a polynomial. In contrast to that result, the construction below requires the function to only be continuously differentiable and gives a *homogeneous* approximating function of degree k. This property is important for our purposes.

Lemma 4.2.1. *Let k be a positive even integer. For any function $V \in \mathcal{H}_k(\mathbb{R}^n)$ and any scalar $\varepsilon > 0$, there exist an even integer r and a homogeneous polynomial p of degree $r + k$ such that*

$$\left\|V(x) - \frac{p(x)}{\|x\|^r}\right\|_{\mathcal{H}} \leq \varepsilon.$$

Proof. Fix $V \in \mathcal{H}_k(\mathbb{R}^n)$ and $\varepsilon > 0$. For every integer m, define the Bernstein polynomial of order m as

$$B_m(x) = \sum_{0\leq j_1,\ldots,j_n\leq m} V\left(\frac{2j_1}{m} - 1, \ldots, \frac{2j_n}{m} - 1\right) \cdot \prod_{s=1}^{n} \binom{m}{j_s} \left(\frac{1+x_s}{2}\right)^{j_s} \left(\frac{1-x_s}{2}\right)^{m-j_s}.$$

The polynomial B_m has degree nm, and has the property that for m large enough, it satisfies

$$\begin{aligned} \sup_{\|x\|\leq 1} |B_m(x) - V(x)| &\leq \frac{\varepsilon}{1+k}, \\ \text{and} \quad \sup_{\|x\|\leq 1} \|\nabla B_m(x) - \nabla V(x)\| &\leq \frac{\varepsilon}{1+k}. \end{aligned} \tag{4.4}$$

See [203, Theorem 4] for a proof. Let m be fixed now and large enough for the above inequalities to hold. Since $V(x)$ is an even function, the function

$$C(x) := \frac{B_m(x) + B_m(-x)}{2}$$

also satisfies (4.4). Because $C(x)$ is even, the function

$$\tilde{C}(x) := \|x\|^k C\left(\frac{x}{\|x\|}\right)$$

is of the form $\frac{p(x)}{\|x\|^r}$, where $p(x)$ is a homogeneous polynomial and r is an even integer. Also, by homogeneity, the degree of $p(x)$ is $r + k$.

It is clear that C and $\tilde{C}$ are equal on the sphere, so

$$\sup_{\|x\|=1} |\tilde{C}(x) - V(x)| \leq \frac{\varepsilon}{1+k}.$$

We argue now that the gradient of $\tilde{C}$ is close to the gradient of V on the sphere. For that, fix $x \in \mathcal{S}^{n-1}$. By Euler's identity for homogeneous functions

$$\langle \nabla \tilde{C}(x), x\rangle - \langle \nabla V(x), x\rangle = k(\tilde{C}(x) - V(x)).$$

Since

$$|\tilde{C}(x) - V(x)| \leq \varepsilon,$$

it is enough to control the part of the gradient that is orthogonal to x. More precisely, let

$$\pi_x(y) := y - \langle x, y\rangle x$$

be the projection of a vector $y \in \mathbb{R}^n$ onto the hyperplane T_x tangent to $\mathcal{S}^{n-1}$ at the point x. The following shows that $\nabla\tilde{C}$ and ∇C are equal when projected on T_x:

$$\begin{aligned}
\pi_x(\nabla \tilde{C}(x)) &= \pi_x\left(k\|x\|^{k-2} C\left(\frac{x}{\|x\|}\right) x\right) \\
&\quad + \pi_x\left(\|x\|^k J\left(\frac{x}{\|x\|}\right)^T \nabla C\left(\frac{x}{\|x\|}\right)\right) \\
&= \pi_x(kC(x)x + (I - xx^T)\nabla C(x)) \\
&= \pi_x(\nabla C(x)).
\end{aligned}$$

Here, the second equation comes from the fact that $\|x\| = 1$ and that the Jacobian of $\frac{x}{\|x\|}$ is equal to $I - xx^T$ on $\mathcal{S}^{n-1}$, and the third equation relies on the fact that the

projection of vector proportional to x onto T_x is zero. Therefore,

$$\begin{aligned}\|\pi_x(\nabla\tilde{C}(x) - \nabla V(x))\| &= \|\pi_x(\nabla C(x) - \nabla V(x))\| \\ &\leq \|\nabla C(x) - \nabla V(x)\| \\ &\leq \frac{\varepsilon}{1+k}.\end{aligned}$$

We conclude by noting that

$$\begin{aligned}\|\nabla\tilde{C}(x) - V(x)\| &\leq \|\pi_x(\nabla\tilde{C}(x) - \nabla V(x))\| \\ &\quad + |\langle x, \nabla\tilde{C}(x) - \nabla V(x)\rangle| \\ &\leq \varepsilon.\end{aligned}$$

□

4.3 Rational Lyapunov Functions

4.3.1 Nonexistence of rational Lyapunov functions

It is natural to wonder whether globally asymptotically stable polynomial vector fields always admit a rational Lyapunov function. We show here that this is not the case, hence also justifying the need for the homogeneity assumption in the statement of our main result (Theorem 4.4.3).

It has been shown in [12] that the polynomial vector field

$$\begin{aligned}\dot{x} &= -x + xy \\ \dot{y} &= -y\end{aligned} \tag{4.5}$$

is globally asymptotically stable (as shown by the Lyapunov function $V(x,y) = \log(1+x^2) + y^2$) but does not admit a polynomial Lyapunov function. We prove here that this vector field does not admit a rational Lyapunov function either. Intuitively, we show that solutions of (4.5) cannot be contained within sublevel sets of rational functions because they can grow exponentially before converging to the origin.

More formally, suppose for the sake of contradiction that the system had a Lyapunov function of the form

$$V(x,y) = \frac{p(x,y)}{q(x,y)},$$

where $p(x,y)$ and $q(x,y)$ are polynomials. Note first that the solution to system (4.5) from any initial condition $(x_0, y_0) \in \mathbb{R}^2$ can be written explicitly:

$$\begin{aligned}x(t) &= x_0 e^{-t} e^{y_0(1-e^{-t})} \\ y(t) &= y_0 e^{-t}.\end{aligned}$$

In particular, a solution that starts from $(x_0, y_0) = (k, \alpha k)$ for $\alpha, k > 1$ reaches the point $(e^{\alpha(k-1)}, \alpha)$ after time

$$t^* = \log(k).$$

As $t^* > 0$, the function V must satisfy

$$V(x(t^*), y(t^*)) < V(x_0, y_0),$$

$$\text{i.e.,} \qquad \frac{p(e^{\alpha(k-1)}, \alpha)}{q(e^{\alpha(k-1)}, \alpha)} < \frac{p(k, \alpha k)}{q(k, \alpha k)}.$$

Fix $\alpha > 1$ and note that since $V(x, \alpha) \to \infty$ as $x \to \infty$, then necessarily the degree of $x \to p(x, \alpha)$ is larger than the degree of $x \to q(x, \alpha)$. We can see from this that the left-hand side of the above inequality grows exponentially in k while the right-hand side grows polynomially, which cannot happen.

4.3.2 Rational Lyapunov functions for homogeneous dynamical systems

We now show that existence of rational Lyapunov functions is necessary for stability of homogeneous vector fields.

Theorem 4.3.1. *Let f be a homogeneous, continuously differentiable function of degree d. Then the system $\dot{x} = f(x)$ is asymptotically stable if and only if it admits a Lyapunov function of the type*

$$V(x) = \frac{p(x)}{(\sum_{i=1}^n x_i^2)^r}, \tag{4.6}$$

where r is a nonnegative integer and p is a homogeneous (positive definite) polynomial of degree $2r + 2$.

Proof. The "if direction" of the theorem is a standard application of Lyapunov's theorem; see e.g. [112, Thm. 4.2].

For the "only if" direction, suppose f is continuously differentiable homogeneous function of degree d, and that the system $\dot{x} = f(x)$ is asymptotically stable. A result of Rosier [179, Thm. 2] (see also [88, Thm. 57.4] [212, Thm. 36] [110, Prop. p.1246]) implies that there exists a function $W \in \mathcal{H}_2(\mathbb{R}^n)$ such that

$$\begin{aligned} W(x) &> 0 \quad \forall x \in \mathcal{S}^{n-1}, \\ -\langle \nabla W(x), f(x) \rangle &> 0 \quad \forall x \in \mathcal{S}^{n-1}. \end{aligned}$$

Since these inequalities are strict and involve continuous functions, we may assume that there exists a $\delta > 0$ such that

$$W(x) \geq \delta \text{ and } -\langle \nabla W(x), f(x) \rangle \geq \delta \; \forall x \in \mathcal{S}^{n-1}.$$

Let

$$f_\infty := \max\{1, \max_{\|x\|=1} \|f(x)\|\}.$$

Lemma 4.2.1 proves the existence of a function $V(x)$ of the form (4.6) that satisfies

$$|V(x) - W(x)| \leq \frac{\delta}{2f_\infty} \qquad \forall x \in \mathcal{S}^{n-1},$$
$$\|\nabla V(x) - \nabla W(x)\| \leq \frac{\delta}{2f_\infty} \qquad \forall x \in \mathcal{S}^{n-1}.$$

Fix $x \in \mathcal{S}^{n-1}$. An application of the Cauchy-Schwarz inequality gives

$$\begin{aligned} |\langle \nabla W(x), f(x)\rangle - \langle \nabla V(x), f(x)\rangle| &\leq \|\nabla W(x) - \nabla V(x)\|\|f(x)\| \\ &\leq \frac{\delta}{2}. \end{aligned}$$

Therefore,

$$V(x) \geq \frac{\delta}{2} \text{ and } -\langle \nabla V(x), f(x)\rangle \geq \frac{\delta}{2} \; \forall x \in \mathcal{S}^{n-1}.$$

□

4.4 An SDP Hierarchy for Searching for Rational Lyapunov Functions

For a rational function of the type in (4.6) to be a Lyapunov function, we need the polynomial V and

$$\begin{aligned} -\dot{V}(x) &:= -\langle \nabla V(x), f(x)\rangle \\ &= \frac{-\|x\|^2\langle \nabla p(x), f(x)\rangle + 2rp(x)\langle x, f(x)\rangle}{\|x\|^{2(r+1)}}, \end{aligned}$$

to be positive definite. This condition is equivalent to the polynomials in the numerators of V and $-\dot{V}$ being positive definite. In this section, we prove a stronger result which shows that there always exists a rational Lyapunov function whose two positivity requirements have "sum of squares certificates". This is valuable because the search over this more restricted class of positive polynomials can be carried out via semidefinite programming while the search over all positive polynomials is NP-hard [156].

Recall that a homogeneous polynomial h of degree $2d$ is a *sum of squares* (sos) if it can be written as $h = \sum_i g_i^2$ for some (homogeneous) polynomials g_i. This is equivalent to existence of a symmetric positive semidefinite matrix Q that satisfies

$$h(x) = m(x)^T Q m(x) \quad \forall x, \tag{4.7}$$

where $m(x)$ is the vector of all monomials of degree d. We say that h is *strictly sos* if it is in the interior of the cone of sos homogeneous polynomials of degree $2d$. This is equivalent to existence of a positive definite matrix Q that satisfies (4.7). Note that a strictly sos homogeneous polynomial is positive definite. We will need the following Positivstellensatz due to Scheiderer.

Lemma 4.4.1 (Scheiderer [184], [185])**.** *For any two positive definite homogeneous polynomials h and g, there exists an integer $q_0 \geq 0$ such that the polynomial hg^q is strictly sos for all integers $q \geq q_0$.*

Theorem 4.4.2. *If a homogeneous polynomial dynamical system admits a rational Lyapunov function of the form*

$$V(x) = \frac{p(x)}{(\sum_i x_i^2)^r},$$

where $p(x)$ is a homogeneous polynomial, then it also admits a rational Lyapunov function

$$W(x) = \frac{\hat{p}(x)}{(\sum_i x_i^2)^{\hat{r}}},$$

where the numerators of W and $-\dot{W}$ are both strictly sos homogeneous polynomials.

Proof. The condition that V be positive definite is equivalent to p being positive definite. The gradient of V is equal to

$$\begin{aligned}\nabla V(x) &= \frac{\|x\|^{2r}\nabla p(x) - 2r\|x\|^{2r-2}p(x)x}{\|x\|^{4r}} \\ &= \frac{\|x\|^2\nabla p(x) - 2rp(x)x}{\|x\|^{2r+2}}.\end{aligned}$$

If we let

$$s(x) := \|x\|^2\nabla p(x) - 2rp(x)x,$$

then the condition that $-\langle \nabla V(x), f(x)\rangle$ be positive definite is equivalent to $-\langle s(x), f(x)\rangle$ being positive definite.

We claim that there exists a positive integer $\hat{q}$, such that

$$W(x) := V^{\hat{q}}(x)$$

satisfies the conditions of the theorem. Indeed, by applying Lemma 4.4.1 with $g = h = p$, there exists q_0, such that p^q is strictly sos for all integers $q \geq q_0$.

Let us now examine the gradient of a function of the type V^q. We have

$$\nabla V^q(x) = qV^{q-1}(x)\nabla V(x) = q\left(\frac{p(x)}{\|x\|^{2r}}\right)^{q-1}\frac{s(x)}{\|x\|^{2r+2}}.$$

Hence,

$$-\langle \nabla V^q(x), f(x)\rangle = \frac{q}{\|x\|^{2rq+2}}p(x)^{q-1}\langle -s(x), f(x)\rangle.$$

Since the homogeneous polynomials

$$p(x) \text{ and } \langle -s(x), f(x)\rangle$$

are both positive definite, by Lemma 4.4.1, there exists an integer q_1 such that

$$p(x)^{q-1}\langle -s(x), f(x)\rangle$$

is strictly sos for all $q \geq q_1$. Taking $\hat{q} = \max\{q_0, q_1\}$ finishes the proof as we can let

$$\hat{p} = p^{\hat{q}}, \hat{r} = r\hat{q}.$$

□

If we denote the degree of $\hat{p}$ by s, then characterization (4.7) of strictly sos homogeneous polynomials applied to the numerator of W and its derivative tells us that there exist positive definite matrices P and Q such that

$$W(x) = \frac{\langle m(x), Pm(x)\rangle}{\|x\|^{2\hat{r}}},$$

and

$$-\dot{W}(x) = \frac{\langle z(x), Qz(x)\rangle}{\|x\|^{2\hat{r}+2}},$$

where $m(x)$ (resp. $z(x)$) here denotes the vector of monomials in x of degree $\frac{s}{2}$ (resp. $\frac{s+d+1}{2}$). Notice that by multiplying W by a positive scalar, we can assume without loss of generality that $P \succeq I$ and $Q \succeq I$.

Putting Theorem 4.3.1 and Theorem 4.4.2 together, we get the main result of this chapter.

Theorem 4.4.3. *A homogeneous polynomial dynamical system $\dot{x} = f(x)$ of degree d is asymptotically stable if and only if there exist a nonnegative integer r, a positive even integer s, with $2r < s$, and symmetric matrices $P \succeq I$ and $Q \succeq I$, such that*

$$\begin{aligned}\langle z(x), Qz(x)\rangle = &-2\|x\|^2\langle J(m(x))^T Pm(x), f(x)\rangle \\ &+ 2rm(x)^T Pm(x)\langle x, f(x)\rangle \quad \forall x \in \mathbb{R}^n,\end{aligned} \tag{4.8}$$

where $m(x)$ (resp. $z(x)$) here denotes the vector of monomials in x of degree $\frac{s}{2}$ (resp. $\frac{s+d+1}{2}$), and $J(m(x))$ denotes the Jacobian of $m(x)$.

For fixed integers s and r with $2r < s$, one can test for existence of matrices $P \succeq I$ and $Q \succeq I$ that satisfy (4.8) by solving a semidefinite program. This gives rise to a hierarchy of semidefinite programs where one tries increasing values of s, and for each s, values of $r \in \{0, \ldots, \frac{s}{2} - 1\}$.

4.5 A Negative Result on Degree Bounds

The sizes of the matrices P and Q that appear in the semidefinite programming hierarchy we just proposed depend on s, but not r. This motivates us to study whether one can bound s as a function of the dimension n and the degree d of the vector field at hand. In this section, we show that the answer to this question is negative. In fact, we prove a stronger result which shows that one cannot bound the degree of the numerator of a rational Lyapunov function based on n and d only (even if one ignores the requirement that the Lyapunov function and its derivative have sos certificates of positivity).

To prove this statement, we build on ideas by Bacciotti and Rosier [178] to construct a family of 2-dimensional degree-3 homogeneous polynomial vector fields that are asymptotically stable but necessitate rational Lyapunov functions whose numerators have arbitrarily high degree.

Proposition 4.5.1. *Let λ be a positive irrational real number and consider the following homogeneous cubic vector field parameterized by the scalar θ:*

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = \begin{pmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{pmatrix}\begin{pmatrix}-2\lambda y(x^2+y^2) - 2y(2x^2+y^2)\\ 4\lambda x(x^2+y^2) + 2x(2x^2+y^2)\end{pmatrix}. \tag{4.9}$$

Then, for $0 < \theta < \pi$, the origin is asymptotically stable. However, for any positive integer s, there exits a scalar $\theta \in (0, \pi)$ such that the vector field in (4.9) does not admit a rational Lyapunov function with a homogeneous polynomial numerator of degree $\leq s$ and a homogeneous polynomial denominator.

The intuition behind this construction is that as $\theta \to 0$, this sequence of vector fields converges to a limit vector field whose trajectories are periodic orbits that cannot be contained within level sets of any rational function. This limit behavior is formalized in the next lemma, which will be used in the proof of the above proposition.

Lemma 4.5.2. *Consider the vector field*

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = f(x,y) = \begin{cases} -2\lambda y(x^2+y^2) - 2y(2x^2+y^2)\\ 4\lambda x(x^2+y^2) + 2x(2x^2+y^2)\end{cases} \tag{4.10}$$

parameterized by a scalar $\lambda > 0$. For all values of λ, the origin is a center[1] of (4.10), but for any irrational value of λ, there exist no two bivariate polynomials p and q such that the rational function

$$W(x,y) := \frac{p(x,y)}{q(x,y)}$$

is nonzero, homogeneous, differentiable, and satisfies

$$\langle \nabla W(x,y), f(x,y)\rangle = 0 \quad \text{for all } (x,y) \in \mathbb{R}^2.$$

[1] By this we mean that all trajectories of (4.10) go on periodic orbits which form closed curves around the origin.

Proof. For the proof of the first claim see [178, Prop.5.2]. Our proof technique for the second claim is also similar to [178, Prop.5.2], except for some minor differences. Suppose for the sake of contradiction that such a function $W(x, y)$ exists. Let k denote the degree of homogeneity of W. We first observe that the function

$$I(x, y) = (x^2 + y^2)(2x^2 + y^2)^\lambda$$

satisfies $\langle \nabla I(x, y), f(x, y) \rangle = 0$. Therefore, on the level set

$$\{(x, y) \in \mathbb{R}^2 \mid I(x, y) = 1\},$$

$W(x, y)$ must be equal to a nonzero constant c. A homogeneity argument shows that

$$W(x, y) = cI(x, y)^{\frac{k}{2(1+\lambda)}} \text{ for all } (x, y) \in \mathbb{R}^2.$$

Hence, by setting $x = 1$,

$$p(1, y) = c(1 + y^2)^{\frac{k}{2(1+\lambda)}}(2 + y^2)^{\frac{k\lambda}{2(1+\lambda)}} q(1, y) \text{ for all } y \in \mathbb{R}. \tag{4.11}$$

Let r be the largest nonnegative integer such that

$$q(1, y) = (1 + y^2)^r \hat{q}(y),$$

where $\hat{q}$ is a univariate polynomial. As a result, $\hat{q}$ must satisfy $\hat{q}(i) \neq 0$, where $i = \sqrt{-1}$. Then, from (4.11), we conclude that

$$p(1, y) = c(1 + y^2)^{r+\frac{k}{2(1+\lambda)}}(2 + y^2)^{\frac{k\lambda}{2(1+\lambda)}} \hat{q}(y) \text{ for all } y \in \mathbb{R}. \tag{4.12}$$

The function $y \to (2 + y^2)^{\frac{k\lambda}{2(1+\lambda)}} \hat{q}(y)$ can be prolonged to a holomorphic function on the open set

$$O_1 := \mathbb{C} \setminus \{y = iv \mid |v| \geq \sqrt{2}\}.$$

Furthermore, since $\hat{q}(i) \neq 0$, there exists an open neighborhood O_2 of i where $\hat{q}$ does not vanish. On the open set $O_1 \cap O_2$, the function

$$y \to (2 + y^2)^{\frac{k\lambda}{2(1+\lambda)}} \hat{q}(y)$$

is holomorphic and does not vanish, and hence by (4.12), the function

$$y \to (1 + y^2)^{r+\frac{k}{2(1+\lambda)}}$$

is also holomorphic on $O_1 \cap O_2$. As a consequence, there exist an integer $\bar{n}$ and a number $\alpha \in \mathbb{C} \setminus \{0\}$ such that

$$\frac{(1 + y^2)^{r+\frac{k}{2(1+\lambda)}}}{(y - i)^{\bar{n}}} \to \alpha$$

as $y \to i$. This implies that

$$r + \frac{k}{2(1+\lambda)} = \bar{n}$$

and contradicts the assumption that λ is an irrational number. □

Proof of Proposition 4.5.1. Consider the positive definite Lyapunov function[2] $V(x,y) = (2x^2 + y^2)^\lambda (x^2 + y^2)$, whose derivative along the trajectories of (4.9) is equal to

$$\dot{V}(x,y) = -\sin(\theta)(2x^2+y^2)^{\lambda-1}(\dot{x}^2+\dot{y}^2).$$

Since $\dot{V}$ is negative definite for $0 < \theta < \pi$, it follows that for θ in this range, the origin of (4.9) is asymptotically stable.

To establish the latter claim of the proposition, suppose for the sake of contradiction that there exists an upper bound $\bar{s}$ such that for all $0 < \theta < \pi$ the system admits a rational Lyapunov function

$$W_\theta(x,y) = \frac{p_\theta(x,y)}{q_\theta(x,y)},$$

where p_θ and q_θ are both homogeneous polynomials and p_θ is of degree at most $\bar{s}$ independently of θ. Note that as a Lyapunov function, W_θ must vanish at the origin, and therefore the degree of q_θ is also bounded by $\bar{s}$. By rescaling, we can assume without loss of generality that the 2-norm of the coefficients of all polynomials p_θ and q_θ is 1.

Let us now consider the sequences $\{p_\theta\}$ and $\{q_\theta\}$ as $\theta \to 0$. These sequences reside in the compact set of bivariate homogeneous polynomials of degree at most $\bar{s}$ with the 2-norm of the coefficients equal to 1. Since every bounded sequence has a converging subsequence, it follows that there must exist a subsequence of $\{p_\theta\}$ (resp. $\{q_\theta\}$) that converges (in the coefficient sense) to some nonzero homogeneous polynomial p_0 (resp. q_0). Define

$$W_0(x,y) := \frac{p_0(x,y)}{q_0(x,y)}.$$

Since convergence of this subsequence also implies convergence of the associated gradient vectors, we get that

$$\dot{W}_0(x,y) = \langle \nabla W_0(x,y), \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} \rangle \leq 0.$$

On the other hand, when $\theta = 0$, the vector field in (4.9) is the same as the one in (4.10) and hence the trajectories starting from any nonzero initial condition go on periodic orbits. This however implies that $\dot{W}_0 = 0$ everywhere and in view of Lemma 4.5.2 we have a contradiction. □

[2] This function is not a polynomial, which can be seen e.g. by noticing that the restriction $V(x,x) = 3^\lambda 2 x^{2(\lambda+1)}$ is not a polynomial.

Remark 7. *It is possible to establish the result of Proposition 4.5.1 without having to use irrational coefficients in the vector field. One approach is to take an irrational number, e.g. π, and then think of a sequence of vector fields given by (4.9) that is parameterized by both θ and λ. We let the k-th vector field in the sequence have $\theta_k = \frac{1}{k}$ and λ_k equal to a rational number representing π up to k decimal digits. Since in the limit as $k \to \infty$ we have $\theta_k \to 0$ and $\lambda_k \to \pi$, it should be clear from the proof of Proposition 4.5.1 that for any integer s, there exists an asymptotically stable bivariate homogeneous cubic vector field with* rational *coefficients that does not have a Lyapunov $\frac{p(x,y)}{q(x,y)}$ where p and q are homogeneous and p has degree less than s.*

4.6 Potential Advantages of Rational Lyapunov Functions over Polynomial Ones

In this section, we show that there are stable polynomial vector fields for which a polynomial Lyapunov function would need to have much higher degree than the sum of the degrees of the numerator and the denominator of a rational Lyapunov function. The reader can also observe that independently of the integer r, the size of the SDP arising form Theorem 4.4.3 that searches for a rational Lyapunov function with a numerator of degree s and a denominator of degree $2r$ is smaller than the size of an SDP that would search for a polynomial Lyapunov function p of degree $s+2$ (by requiring p and $-\dot{p}$ to be sums of squares), even when p is taken to be homogeneous. Therefore, for some vector fields, a search for a rational Lyapunov function instead of a polynomial one can be advantageous.

Proposition 4.6.1. *Consider the following homogeneous polynomial vector field parameterized by the scalar θ:*

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} = f_\theta(x,y) = 2R(\theta)\begin{pmatrix} x\left(x^4+2\,x^2\,y^2-y^4\right)\\ y\left(-x^4+2\,x^2\,y^2+y^4\right)\end{pmatrix}, \tag{4.13}$$

where

$$R(\theta) := \begin{pmatrix}-\sin(\theta) & -\cos(\theta)\\ \cos(\theta) & -\sin(\theta)\end{pmatrix}.$$

Then, for $\theta \in (0,\pi)$, the vector field f_θ admits the following rational Lyapunov function

$$W(x,y) = \frac{x^4+y^4}{x^2+y^2}$$

and hence is asymptotically stable. However, for any positive integer $\bar{s}$, there exits a scalar $\theta \in (0,\pi)$ such that f_θ does not admit a polynomial Lyapunov function of degree $\leq \bar{s}$.

Once again, the intuition is that as $\theta \to 0$, f_θ converges to a vector field f_0 whose trajectories are periodic orbits. This time however, these orbits will exactly traverse the level sets of the rational function W and cannot be contained within level sets of any polynomial. Our proof will utilize the following independent lemma about univariate polynomials.

Lemma 4.6.2. *There exist no two univariate polynomials $\tilde{p}$ and $\tilde{q}$, with $\tilde{q}$ non-constant, that satisfy*

$$\tilde{p}(x^2) = \tilde{q}\left(\frac{x^4+1}{x^2+1}\right) \quad \forall x \in \mathbb{R}.$$

Proof. Assume for the sake of contradiction that such polynomials exist. For every nonnegative scalar u, there exists a scalar x such that $u = x^2$. Therefore,

$$\tilde{p}(u) = \tilde{q}\left(\frac{u^2+1}{u+1}\right) \quad \forall u \geq 0.$$

The expression above is an equality between two univariate rational functions valid on $[0, \infty)$. Since both rational functions are well-defined on $(-1, \infty]$, the equality holds on that interval as well:

$$\tilde{p}(u) = \tilde{q}\left(\frac{u^2+1}{u+1}\right) \quad \forall u > -1.$$

We get a contradiction by taking $u \to -1$ as the left hand side converges to $\tilde{p}(-1)$, while the right hand side diverges to ∞. □

Proof of Proposition 4.6.1. Let us first prove that W is a rational Lyapunov function associated with the vector field f_θ whenever $\theta \in (0, \pi)$. It is clear that W is positive definite and radially unbounded. A straightforward calculation shows that

$$f_\theta(x, y) = R(\theta)\ (x^2 + y^2)^2\ \nabla W(x, y).$$

Hence,

$$\begin{aligned} -\dot{W}(x, y) &= -\langle \nabla W(x, y), f_\theta(x, y) \rangle \\ &= -(x^2 + y^2)^2 \nabla W(x, y)^T R(\theta) \nabla W(x, y) \\ &= \sin(\theta)(x^2 + y^2)^2 \|\nabla W(x, y)\|^2. \end{aligned}$$

Note that the function $\|\nabla W\|^2$ is positive definite as

$$W(x, y) = \frac{1}{2}\langle \begin{pmatrix} x \\ y \end{pmatrix}, \nabla W(x, y) \rangle$$

and W is positive definite. This proves that when $0 < \theta < \pi$, the vector field f_θ is asymptotically stable with W as a Lyapunov function.

To prove the latter claim of the proposition, suppose for the sake of contradiction that there exists an upper bound $\bar{s}$ such that for all $0 < \theta < \pi$ the system admits a polynomial Lyapunov function of degree at most $\bar{s}$. By an argument similar to that in the proof of Proposition 4.5.1, there must exist some nonzero polynomial p_0, with $p_0(0) = 0$, that satisfies

$$\dot{p}_0(x, y) := \langle \nabla p_0(x, y), f_0(x, y) \rangle \leq 0 \quad \forall (x, y) \in \mathbb{R}^2.$$

We claim that p_0 must be constant on the level sets of W. To prove that, consider an arbitrary positive scalar γ and the level set

$$M_\gamma := \{(x, y) \in \mathbb{R}^2 \mid W(x, y) = \gamma\}.$$

Since W is homogeneous and positive definite, M_γ is closed and bounded. In addition, f_0 is continuously differentiable and does not vanish on M_γ. Moreover, trajectories starting in M_γ remain in M_γ as

$$\langle \nabla W(x, y), f_0(x, y) \rangle = \sin(0)(x^2 + y^2)^2 \|\nabla W(x, y)\|^2 = 0.$$

Hence, by the Poincaré-Bendixson Criterion [112, Lem 2.1], the set M contains a periodic orbit of f_0.

Let $z_1, z_2 \in M_\gamma$. We know that the trajectory starting from z_1 must visit z_2. Since $\dot{p}_0 \leq 0$, we must have $p_0(z_1) \leq p_0(z_2)$. Similarly, we must also have $p_0(z_2) \leq p_0(z_1)$, and therefore

$$p_0(z_1) = p_0(z_2).$$

Since we now know that p_0 is constant on the level sets of W, there must exist a function $g : \mathbb{R} \to \mathbb{R}$ such that

$$p_0(x, y) = g(W(x, y)) = g\left(\frac{x^4 + y^4}{x^2 + y^2}\right).$$

This proves that

$$p_0(x, y) = p_0(x, -y) = p_0(-x, y) = p_0(-x, -y).$$

Therefore, there exists a polynomial p such that

$$p_0(x, y) = p(x^2, y^2) = g\left(\frac{x^4 + y^4}{x^2 + y^2}\right). \tag{4.14}$$

Setting $y = 0$, we get that $p(x^2, 0) = g(x^2)$. Hence, $p(u, 0) = g(u)$ for all $u \geq 0$. Taking

$$u = \frac{x^4 + y^4}{x^2 + y^2},$$

the second equality in (4.14) gives

$$p(x^2, y^2) = p\left(\frac{x^4 + y^4}{x^2 + y^2}, 0\right).$$

Setting $y = 1$, we get that the polynomial p satisfies

$$p(x^2, 1) = p\left(\frac{x^4 + 1}{x^2 + 1}, 0\right).$$

If we let $\tilde{p}(x) := p(x,1)$ and $\tilde{q}(x) := p(x,0)$, then in view of Lemma 4.6.2 and the fact that $\tilde{q}$ is not constant, we have a contradiction. □

Example 2. *Consider the vector field f_θ in (4.13) with $\theta = 0.05$. One typical trajectory of this vector field is depicted in Figure 4.1. We use the modeling language YALMIP [131] and the SDP solver mosek [22] to search for rational and polynomial Lyapunov functions for this vector field.*

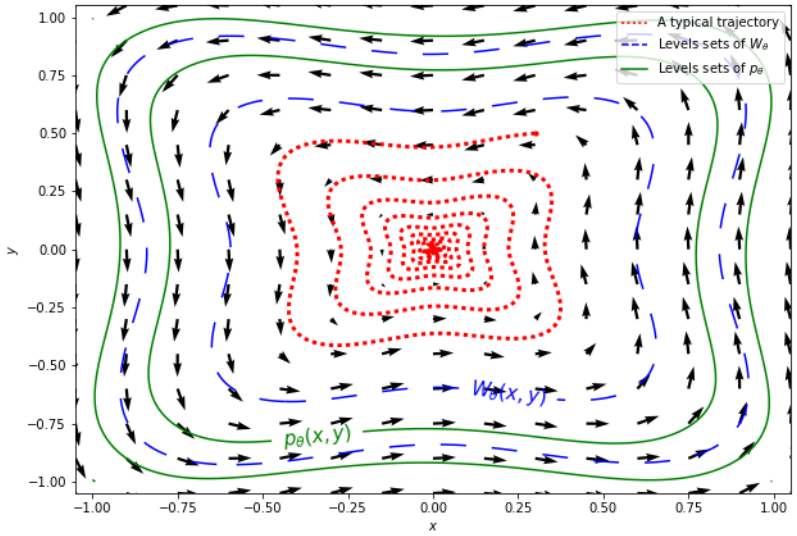


Figure 4.1: A typical trajectory of the vector field f_θ in (4.13) with $\theta = 0.05$, together with the level sets of the Lyapunov functions W_θ and p_θ.

We know that for $\theta = 0.05$, the vector field is asymptotically stable. Therefore, by Theorem 4.4.3, the semidefinite programming hierarchy described in Section 4.4 is guaranteed to find a rational Lyapunov function. The first round to succeed corresponds to $(s,r) = (4,1)$, and produces the feasible solution

$$W_\theta(x,y) = \frac{16.56x^4 + 16.56y^4 + 0.04x^2y^20.17x^3y - 0.17xy^3}{x^2+y^2}$$

If we look instead for a polynomial Lyapunov function, i.e. $r = 0$, the lowest degree for which the underlying SDP is feasible corresponds to $s = 8$. The Lyapunov function that our solver returns is the following polynomial:

$$\begin{aligned} p_\theta(x,y) = {} & 42.31x^8 + 42.31y^8 + 6.5xy^7 - 6.5x^7y - 100.94x^2y^6 \\ & -100.94x^6y^2 + 19.86x^5y^3 - 19.86x^3y^5 + 166.65x^4y^4. \end{aligned}$$

As all bivariate nonnegative homogeneous polynomials are sums of squares, infeasibility of our SDP for $s = 2,4,6$ means that f_θ admits no homogeneous polynomial Lyapunov function of degree lower than 8. Two level sets of W_θ and p_θ are shown in Figure 4.1 and they look quite similar.

4.7 Conclusions and Future Directions

We showed in this chapter that existence of a rational Lyapunov function is necessary and sufficient for asymptotic stability of homogeneous continuously differentiable vector fields. In the case where the vector field is polynomial, we constructed an SDP hiearachy that is guaranteed to find this Lyapunov function. The number of variables and constraints in this SDP hiearachy depend only on s, the degree of the numerator of the candidate Lyapunov function, and not on r, the degree of its denominator. To our knowledge, this theorem constitutes one of the few results in the theory of nonlinear dynamical systems which guarantees existence of algebraic certificates of stability that can be found by convex optimization (in fact, the only one we know of which applies to polynomial vector fields that are not exponentially stable). Regarding degree bounds, we proved that even for homogeneous polynomial vector fields of degree 3 on the plane, the degree s of the numerator of such a rational Lyapunov function might need to be arbitrarily high. We also gave a family of homogeneous polynomial vector fields of degree 5 on the plane that all share a simple low-degree rational Lyapunov function, but require polynomial Lyapunov functions of arbitrarily high degree. Therefore, there are asymptotically stable polynomial vector fields for which a search for a rational Lyapunov function is much cheaper than a search for a polynomial one. We leave the following two questions for future research:

- Can r be upperbounded by a computable function of the coefficients of the vector field f? In particular, can r always be taken to be zero? Or equivalently, do asymptotically stable homogeneous vector fields always admit a homogeneous polynomial Lyapunov function?

- Similarily, can s be upperbounded as a computable function of the coefficients of the vector field f? We have shown that s cannot be upperbounded by a function of the dimension n and the degree d of the vector field only.

Finally, while our focus in this chapter was on analysis problems, we hope that our work also motivates further research on understanding the power and limitations of rational Lyapunov functions for *controller design* problems.

Chapter 5

A Globally Asymptotically Stable Polynomial Vector Field with Rational Coefficients and no Local Polynomial Lyapunov Function

5.1 Introduction and Motivation

We are concerned in this chapter with a continuous time dynamical system

$$\dot{\mathbf{x}} = f(\mathbf{x}), \tag{5.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a *polynomial* and has an equilibrium point at the origin, i.e., $f(0) = 0$. Polynomial differential equations appear throughout engineering and the sciences and the study of stability of their equilibrium points has been a problem of long-standing interest to mathematicians and control theorists.

We recall that the origin of (5.1) is said to be a *locally asymptotically stable* (LAS) equilibrium if it is *stable in the sense of Lyapunov* (i.e., if for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ such that $\|\mathbf{x}(0)\| < \delta \;\Rightarrow \|\mathbf{x}(t)\| < \epsilon$ for all $t \geq 0$) and if there exists a scalar $\hat{\delta} > 0$ such that

$$\|\mathbf{x}(0)\| < \hat{\delta} \;\Rightarrow\; \lim_{t\to\infty} \mathbf{x}(t) = 0.$$

We say that the origin of (5.1) is a *globally asymptotically stable* (GAS) if it is stable in the sense of Lyapunov and if $\lim_{t\to\infty} \mathbf{x}(t) = 0$ for any initial condition $x(0)$ in $\mathbb{R}^n$.

We also recall (see, e.g., [112]) that the origin of (5.1) is LAS if there exists a continuously differentiable (Lyapunov) function $V : \mathbb{R}^n \to \mathbb{R}$ that vanishes at the origin and satisfies $V(\mathbf{x}) > 0$ and $-\langle \nabla V(\mathbf{x}), f(\mathbf{x})\rangle > 0$ for all $\mathbf{x} \in S \setminus \{0\}$, where S is a neighborhood of the origin. Moreover, if V is in addition radially unbounded (i.e., satisfies $V(\mathbf{x}) \to \infty$ when $\|\mathbf{x}\| \to \infty$) and if $S = \mathbb{R}^n$, then the origin is GAS. We call a function satisfying the former (resp. the latter) requirements a *local (resp. global) Lyapunov function*. It is also well known that existence of such Lyapunov functions is not only sufficient, but also necessary for local/global asymptotic stability [112].

Since the vector field in (5.1) is polynomial, it is natural to search for Lyaponuv functions that are polynomials themselves. This approach has become widely popular in the last couple of decades due to the advent of optimization-based algorithms that automate the search for a polynomial Lyapunov function. Arguably, the most prominent such algorithm is based on *sum of squares optimization,* which reduces this search to a semidefinite program [156, 149, 95, 102, 53, 94, 211]. Alternatives to this approach that are based on linear programming or other algebraic techniques have also appeared in recent years [107, 8, 108, 37]. As the algorithmic construction of polynomial Lyapunov functions has been the focus of intense research in recent years, it is natural to ask whether existence of a Lyapunov function within this class is guaranteed. This is the case, e.g., if the goal is to prove exponential stability of an equilibrium point over a bounded region [160], [161]. Our focus in this chapter, however, is on the basic question of whether asymptotic stability of an equilibrium point implies existence of a polynomial Lyapunov function. As is well known, the answer is positive when the degree of the vector field in (5.1) is equal to one. Indeed, asymptotically stable linear systems always admit a quadratic Lyapunov function.

Unlike the linear case, stable polynomial vector fields of degree as low as 2 may fail to admit a polynomial Lyapunov function. Indeed, in [12], it is shown that the simple vector field

$$\begin{aligned} \dot{x} &= -x + xy \\ \dot{y} &= -y \end{aligned} \tag{5.2}$$

is globally asymptotically stable (e.g. as certified by the Lyapunov function $V(x,y) = \log(1+x^2)+y^2$), but does not admit a (global) polynomial Lyapunov function. Note however, that the linearization of (5.2) around the origin is asymptotically stable, and hence this nonlinear system admits a local quadratic Lyapunov function.

In [29, Prop. 5.2], Bacciotti and Rosier show that the vector field

$$\begin{aligned} \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = &\begin{pmatrix} -2\lambda y(x^2+y^2) - 2y(2x^2+y^2) \\ 4\lambda x(x^2+y^2) + 2x(2x^2+y^2) \end{pmatrix} \\ &- (x^2+y^2)\begin{pmatrix} 4\lambda x(x^2+y^2) + 2x(2x^2+y^2) \\ 2\lambda y(x^2+y^2) - 2y(2x^2+y^2) \end{pmatrix} \end{aligned} \tag{5.3}$$

is globally asymptotically stable for any scalar $\lambda \geq 0$ (e.g. as certified by the Lyapunov function $V_\lambda(x,y) = (x^2+y^2)(2x^2+y^2)^\lambda$), but does not admit a *local* polynomial Lyapunov function for any λ which is *irrational.*[1] However, the validity of this statement crucially relies on the parameter λ being irrational. Indeed, for any rational value of $\lambda \geq 0$, the system admits a global polynomial Lyaponuv function, which is e.g. simply an appropriate integer power of V_λ.

Our contribution in this chapter is to give an example of a (globally) asymptotically stable polynomial vector field with *rational coefficients* that does not admit a *local* polynomial (or even analytic) Lyaponuv function. Our construction is inspired by and is similar to that of Bacciotti and Rosier [29]. However, by adapting their

[1] In fact, they show that for irrational λ, the system (5.3) does not even admit a local analytic Lyaponuv function.

underlying proof technique, we are able to prove stability with a Lyapunov function which is the ratio of two polynomials. This allows us to use only rational coefficients in the construction of the vector field.[2]

Our interest in studying polynomial vector fields with rational coefficients partly stems from the fact that in practice, most (if not all) vector fields that are analyzed on a computer (e.g. by an optimization-based algorithm) have rational coefficients. Therefore, if it was true that such vector fields always had polynomial Lyapunov functions, one could restrict attention to this function class for all practical purposes and use techniques such as sum of squares optimization to algorithmically find these Lyapunov functions. Because of this practical motivation, existence of the counterexample that we present in this chapter was regarded as a significant unresolved question in the community; see e.g. the ending paragraph in [129, Sect. IV].

Polynomial vector fields with rational coefficients are also important from the viewpoint of complexity analysis in the standard Turing model. For example, it is not known whether the problem of testing local asymptotic stability is decidable for this class of vector fields. Indeed, this is an outstanding open problem suggested by Arnold, which appears e.g. in [69], [25]:

> "Let a vector field be given by polynomials of a fixed degree, with rational coefficients. Does an algorithm exist, allowing to decide, whether the stationary point is stable?"

In [69], Arnold is quoted to have conjectured that the answer to the above question is negative:

> "My conjecture has always been that there is no algorithm for some sufficiently high degree and dimension."

This conjecture also motivates the example in this chapter: if it was true that LAS polynomial vector fields with rational coefficients always admitted polynomial Lyaponuv functions of a computable degree, then the problem of testing stability would become decidable. This is because one can e.g. use the quantifier elimination theory of Tarski and Seidenberg [196], [187] to test, in finite time, whether a polynomial vector field admits a local polynomial Lyaponuv function of a given degree.

We end our introduction by noting that, interestingly, there is a parallel to these questions in the study of switched linear systems in discrete time. There, the problem of testing asymptotic stability is similarly not known to be decidable [47, Problem 10.2], [105]. One can show, however, that if the so called "finiteness conjecture" [120] is true for rational matrices, then asymptotic stability becomes decidable. This conjecture is known to be false over the reals [91], but is currently unresolved for rational matrices [106].

[2]Note that by rescaling, one can always change a polynomial vector field with rational coefficients to a polynomial vector field with integer coefficients without changing the properties of stability or validity of a candidate Lyapunov function.

5.2 The Main Result

Our contribution in this chapter is to prove the following theorem.

Theorem 5.2.1. *The polynomial vector field*

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = f(x, y), \tag{5.4}$$

with

$$f(x,y) = \begin{pmatrix} -2y(-x^4 + 2x^2y^2 + y^4) \\ 2x(x^4 + 2x^2y^2 - y^4) \end{pmatrix} - (x^2 + y^2) \begin{pmatrix} 2x(x^4 + 2x^2y^2 - y^4) \\ 2y(-x^4 + 2x^2y^2 + y^4) \end{pmatrix},$$

is globally asymptotically stable but does not admit an analytic Lyapunov function even locally.

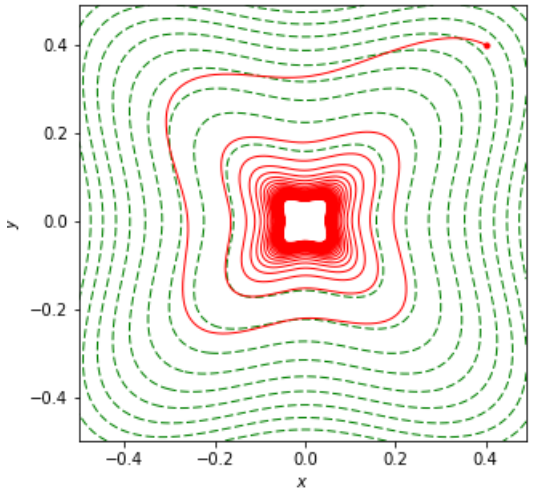


Figure 5.1: A typical trajectory of the vector field in (5.4) and the level sets of the Lyapunov function W in (5.5).

Proof. We prove that the vector field in (5.4) is globally asymptotically stable by means of the rational Lyapunov function defined as

$$W(x,y) = \frac{x^4 + y^4}{x^2 + y^2} \; \forall (x,y) \neq (0,0), \text{ and } W(0,0) = 0. \tag{5.5}$$

Note that the function W is continuously differentiable on $\mathbb{R}^2$, positive definite (i.e., satisfies $W(x,y) > 0$ for all $(x,y) \neq (0,0)$), and radially unbounded. Radial unboundedness can be seen, e.g., by noting that since $||(x,y)^T||_2 \leq 2^{1/4}||(x,y)^T||_4$ for

all $(x, y) \in \mathbb{R}^2$, we have

$$W(x,y) = \frac{||(x,y)^T||_4^4}{||(x,y)^T||_2^2} \geq \frac{1}{2}||(x,y)^T||_2^2, \ \forall (x,y) \in \mathbb{R}^2.$$

Let us examine the gradient of W. A straightforward calculation gives

$$\nabla W(x,y) = \frac{1}{(x^2+y^2)^2} \begin{pmatrix} a(x,y) \\ b(x,y) \end{pmatrix},$$

where $a(x,y) = 2x(x^4 + 2x^2y^2 - y^4)$ and $b(x,y) = 2y(-x^4 + 2x^2y^2 + y^4)$.

If we let $f_0 = \begin{pmatrix} -b \\ a \end{pmatrix}$, and $f_1 = -(x^2+y^2)\begin{pmatrix} a \\ b \end{pmatrix}$, then $f = f_0 + f_1$, and

$$\begin{aligned} \langle \nabla W, f \rangle &= \langle \nabla W, f_0 \rangle + \langle \nabla W, f_1 \rangle \\ &= 0 - \frac{a^2+b^2}{x^2+y^2}. \end{aligned}$$

We show that $\langle \nabla W, f \rangle$ is negative when $(x,y) \neq (0,0)$ by observing that for every $(x,y) \in \mathbb{R}^2 \setminus \{(0,0)\}$, $a(x,y)$ and $b(x,y)$ cannot both be zero. Indeed, if $a(x,y) = b(x,y) = 0$ for some $(x,y) \in \mathbb{R}^2$, then

$$ya(x,y) + xb(x,y) = 8(xy)^3 = 0,$$

therefore $x = 0$ or $y = 0$. If $x = 0$ for example (the case $y = 0$ is similar), then $b(x,y) = 2y^5$, and hence $y = 0$ as well. This shows that

$$\langle \nabla W(x,y), f(x,y) \rangle < 0 \quad \forall (x,y) \neq (0,0),$$

and hence W is a global Lyapunov function which proves that the vector field is GAS.

Let us now show that f does not admit an analytic Lyapunov function locally. Assume for the sake of contradiction that such a function $p : \mathbb{R}^2 \to \mathbb{R}$ exists. By analyticity, $p = \sum_{k=0}^{\infty} p_k$, where p_k is a homogeneous polynomial of degree k. Let p_{k_0} be the first non-vanishing term. Note that $k_0 \geq 2$ as

$$p(0,0) = 0 \Rightarrow p_0 = 0,$$

$$p \geq 0, p(0,0) = 0 \Rightarrow \nabla p(0,0) = (0,0)^T \Rightarrow p_1(x,y) = 0, \forall (x,y) \in \mathbb{R}^2.$$

Here, the first implication follows from the fact that the origin is a global minimum for p. Observe now that

$$\begin{aligned} \langle \nabla p, f \rangle &= \langle \nabla \sum_{k=k_0}^{\infty} p_k, f_0 + f_1 \rangle \\ &= \langle \nabla p_{k_0}, f_0 \rangle + q, \end{aligned}$$

where $q := \langle \nabla p_{k_0}, f_1 \rangle + \sum_{k=k_0+1}^{\infty} \langle \nabla p_k, f_0 + f_1 \rangle$. Note that all terms in q have degree higher than the degree of the (homogeneous) polynomial $\langle \nabla p_{k_0}, f_0 \rangle$. This is because f_1 has higher degree than f_0 and the index of the sum in the definition of q starts at k_0+1. Since $\langle \nabla p, f \rangle \leq 0$ (as we are assuming that p is a Lyapunov function), and since $\langle \nabla p_{k_0}, f_0 \rangle$ constitutes the terms of $\langle \nabla p, f \rangle$ of lowest order, it must be that $\langle \nabla p_{k_0}, f_0 \rangle$ is nonpositive in a small enough neighborhood of the origin. But as $\langle \nabla p_{k_0}, f_0 \rangle$ is homogeneous, this implies that

$$\langle \nabla p_{k_0}(x, y), f_0(x, y) \rangle \leq 0 \quad \forall (x, y) \in \mathbb{R}^2. \tag{5.6}$$

We now claim that the (homogeneous) polynomial p_{k_0} must be constant on the 1-level set of W, which we denote by

$$M := \{(x, y) \in \mathbb{R}^2 \mid W(x, y) = 1\}.$$

Since W is continuous (resp. radially unbounded), it follows that M is closed (resp. bounded). In addition, f_0 is continuously differentiable and does not vanish on M, as we have already argued that $a(x, y)$ and $b(x, y)$ cannot simultaneously vanish except at the origin. Moreover, trajectories of the vector field f_0 that start in M remain in M as one can verify that

$$\langle \nabla W, f_0 \rangle = 0.$$

Hence, by the Poincaré-Bendixson Criterion (see e.g. [112, Lemma 2.1]), the set M contains a periodic orbit of f_0.

Since M is a one-dimensional connected manifold, the trajectory of f_0 starting from a point $\mathbf{z_0} \in M$ on this periodic orbit can only return to $\mathbf{z_0}$ by traversing all points in M. Hence, the periodic orbit coincides with M. In view of the fact that $\langle \nabla p_{k_0}, f_0 \rangle \leq 0$ as established in (

Note that the constant c must be nonzero or else, by homogeneity, the polynomial p_{k_0} would be identically zero, contradicting the definition of k_0. As a consequence,

$$p_{k_0} \text{ and } cW^{\frac{k_0}{2}}$$

are two nonzero homogeneous functions of degree k_0 that are equal on M. Since M intersects all the lines passing through the origin, and since any homogeneous function $u : \mathbb{R}^2 \to \mathbb{R}$ of degree k_0 satisfies $u(\lambda x, \lambda y) = \lambda^{k_0} u(x, y)$ for all $\lambda \in \mathbb{R}$ and all $(x, y) \in \mathbb{R}^2$, we get that

$$p_{k_0}(x, y) = cW^{\frac{k_0}{2}}(x, y) \quad \forall (x, y) \in \mathbb{R}^2.$$

This implies the following polynomial identity

$$(x^2 + y^2)^{k_0} p_{k_0}^2(x, y) = c^2 (x^4 + y^4)^{k_0},$$

which gives a contradiction as $(x, y) = (\sqrt{-1}, 1)$ makes only the left-hand side vanish. □

The vector field in (5.4) is a polynomial of degree 7 in two variables. We leave open the problem of determining the minimum degree of a polynomial vector field with rational coefficients for which the statement of Theorem 5.2.1 holds. Note also that although the vector field in (5.4) does not admit a polynomial Lyapunov function, it admits a rational one (i.e., a ratio of two polynomials). We leave the question of determining whether LAS polynomial vector fields with rational coefficients admit a local rational Lyapunov function for future research. We have recently shown in [4] that one cannot hope for a *global* rational Lyapunov function in general. On the other hand, [4] also shows that if the vector field in question is homogeneous, then asymptotic stability implies existence of a rational Lyaponuv function.

Chapter 6

Learning Dynamical Systems with Side Information

6.1 Motivation and problem formulation

In several safety-critical applications, one has to learn the behavior of an unknown dynamical system from noisy observations of a very limited number of trajectories. For example, to autonomously land an airplane that has just gone through engine failure, limited time is available to learn the modified dynamics of the plane before appropriate control action can be taken. Similarly, when a new infectious disease breaks out, few observations are initially available to understand the dynamics of contagion. In situations of this type where data is limited, it is essential to exploit "side information"—e.g. physical laws or contextual knowledge—to assist the task of learning.

More formally, our interest in this paper is to learn a continuous-time dynamical system of the form

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)), \tag{6.1}$$

over a given compact set $\Omega \subset \mathbb{R}^n$ from noisy observations of a limited number of its trajectories. Here, $\dot{x}(t)$ denotes the time derivative of the state $x(t) \in \mathbb{R}^n$ at time t. We assume that the unknown vector field f that is to be learned is continuously differentiable over an open set containing Ω, an assumption that is often met in applications. In our setting, we have access to a training set of the form

$$\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i), \quad i = 1, \ldots, N\}, \tag{6.2}$$

where $\mathbf{x}_i \in \Omega$ (resp. $\mathbf{y}_i \in \mathbb{R}^n$) is a possibly noisy measurement of the state of the dynamical system (resp. of $f(x_i)$). Typically, this training set is obtained from observation of a few trajectories of (6.1). The vectors y_i could be either directly accessible (e.g., from sensor measurements) or approximated from the state variables using a finite-difference scheme.

Finding a vector field $f_{\mathcal{F}}$ that best agrees with the training set $\mathcal{D}$ among a particular class $\mathcal{F}$ of continuously-differentiable functions amounts to solving the opti-

mization problem

$$f_{\mathcal{F}} \in \arg\min_{p \in \mathcal{F}} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} \ell(p(\mathbf{x}_i), \mathbf{y}_i), \tag{6.3}$$

where $\ell(\cdot, \cdot)$ is some loss function that penalizes deviation of $p(x_i)$ from y_i. For instance, $\ell(\cdot, \cdot)$ could simply be the ℓ_2 loss function

$$\ell_2(u, v) := \|u - v\|_2 \quad \forall u, v \in \mathbb{R}^n,$$

though the computational machinery that we propose can readily handle various other convex loss functions (see Section 6.3).

In addition to fitting the training set $\mathcal{D}$, we desire for our learned vector field $f_{\mathcal{F}}$ to generalize well, i.e., to be consistent as much as possible with the behavior of the unknown vector field f on all of Ω. Indeed, the optimization problem in (6.3) only dictates how the candidate vector field should behave on the training data. This could easily lead to overfitting, especially if the function class $\mathcal{F}$ is large and the observations are limited. Let us demonstrate this phenomenon by a simple example.

Example 3. *Consider the two-dimensional vector field*

$$f(x_1, x_2) := (-x_2, x_1)^T. \tag{6.4}$$

The trajectories of the system $\dot{\mathbf{x}}(t) = f(\mathbf{x}(t))$ from any initial condition are given by circular orbits. In particular, if started from the initial condition $x_{init} = (1, 0)^T$, the trajectory is given by $\mathbf{x}(t, x_{init}) = (\cos(t), \sin(t))^T$. Hence, for any function $g : \mathbb{R}^2 \to \mathbb{R}^2$, the vector field

$$h(\mathbf{x}) := f(\mathbf{x}) + (x_1^2 + x_2^2 - 1)g(\mathbf{x}) \tag{6.5}$$

agrees with f on the sample trajectory $\mathbf{x}(t, x_{init})$. However, the behavior of the trajectories of h depends on the arbitrary choice of the function g. If $g(\mathbf{x}) = \mathbf{x}$ for instance, the trajectories of h starting outside of the unit disk diverge to infinity. See fig. 6.1 for an illustration.

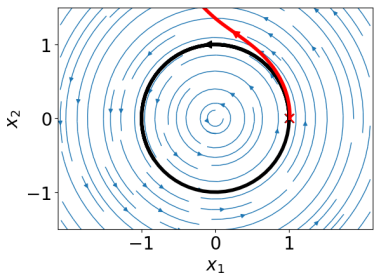


Figure 6.1: Streamplot of the vector field f in (6.4) (in blue), together with two sample trajectories of the vector field h in (6.5) with $g(x) = x$ when started from $(1, 0)^T$ (drawn in black) and from $(1.01, 0)^T$ (drawn in red). The trajectories of f and h match exactly when started from $(1, 0)^T$, but get arbitrarily far from each other when started from $(1.01, 0)^T$.

To address the issue of insufficiency of data and to avoid overfitting, we would like to exploit the fact that in many applications, one may have contextual information about the vector field f without knowing f precisely. We call such contextual information *side information*. Formally, every side information is a subset S of the set

of all continuously-differentiable functions that the vector field f is known to belong to. Equipped with a list of side information $S_1, \ldots, S_k$, our goal is to replace the optimization problem in (6.3) with

$$f_{\mathcal{F}\cap S_1\cap\cdots\cap S_k} \in \operatorname*{arg\,min}_{p\in\mathcal{F}\cap S_1\cap\cdots\cap S_k} \sum_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}} \ell(p(\mathbf{x}_i),\mathbf{y}_i), \tag{6.6}$$

i.e., to find a vector field $f_{\mathcal{F}\cap S_1\cap\cdots\cap S_k} \in \mathcal{F}$ that is closest to f on the training set $\mathcal{D}$ and also satisfies the side information $S_1, \ldots, S_k$ that f is known to satisfy.

6.1.1 Outline and contributions of the paper

In the remainder of this paper, we build on the mathematical formalism we have introduced thus far and make problem (6.6) more concrete and amenable to computation. In Section 6.2, we identify six notions of side information that are commonly encountered in practice and that have attractive convexity properties, therefore leading to a convex optimization formulation of problem (6.6). In Section 6.3, we show that when the function class $\mathcal{F}$ is chosen as the set of polynomial functions of a given degree, then any combination of our six notions of side information can be enforced by semidefinite programming. The derivation of these semidefinite programs leverages ideas from sum of squares optimization, a concept that we briefly review in the same section for the convenience of the reader. In Section 6.4, we demonstrate the applicability of our approach on three examples from epidemiology, classical mechanics, and cell biology. In each example, we show how common sense and contextual knowledge translate to the notions of side information we present in this paper. Furthermore, in each case, we show that by imposing side information via semidefinite programming, we can learn the behavior of the unknown dynamics from a very limited set of observations. In our epidemiology example, we also show the benefits of our approach for a downstream task of optimal control (Section 6.4.4). In Section 6.5, we study the question of how well trajectories of a continuously differentiable vector field that satisfies some side information can be approximated by trajectories of a polynomial vector field that satisfies the same side information either exactly or approximately. We end the paper with a discussion of future research directions in Section 6.6.

6.1.2 Related work

The idea of using sum of squares and semidefinite optimization for verifying various properties of a known dynamical system has been the focus of much research in the control and optimization communities [156, 46, 124, 49]. Our work borrows some of these techniques to instead impose a desired set of properties on a candidate dynamical system that is to be learned from data.

Learning dynamical systems from data is an important problem in the field of system identification. Various classes of vector fields have been proposed throughout the years as candidates for the function class $\mathcal{F}$ in (6.3); e.g., reproducing kernel Hilbert spaces [189, 190, 52], Guassian mixture models [113], and neural networks

[209, 80]. Some recent approaches to learning dynamical systems from data impose additional properties on the candidate vector field. These properties include contraction [189, 74], stabilizability [191], and stability [119], and can be thought of as side information. In contrast to our work, imposing these properties requires formulation of nonconvex optimization problems, which can be hard to solve to global optimality. Furthermore, these references impose the desired properties only on sample trajectories (as opposed to the entire space where the properties are known to hold), or introduce an additional layer of nonconvexity to impose the constraints globally.

We also note that the problem of fitting a polynomial vector field to data has appeared e.g. in [183], though the focus there is on imposing sparsity of the coefficients of the vector field as opposed to side information. The closest work in the literature to our work is that of Hall on shape-constrained regression [89, Chapter 8], where similar algebraic techniques are used to impose constraints such as convexity and monotonicity on a polynomial regressor. See also [63] for some statistical properties of these regressors and several applications. Our work can be seen as an extension of this approach to a dynamical system setting.

6.2 Side information

In this section, we identify six types of side information which we believe are useful in practice (see, e.g., Section 6.4) and that lead to a convex formulation of problem (6.6). For example, we will see in Section 6.3 that semidefinite programming can be used to impose any list of side information constraints of the six types below on a candidate vector field that is parameterized as a polynomial function. The set Ω that appears in these definitions is a compact subset of $\mathbb{R}^n$ over which we would like to learn an unknown vector field f. Throughout this paper, the notation $f \in C_1^\circ(\Omega)$ denotes that f is continuously differentiable over an open set containing Ω.

- **Interpolation at a finite set of points.** For a set of points $\{(\mathbf{x}_i, \mathbf{y}_i) \in \Omega \times \mathbb{R}^n\}_{i=1}^m$, we denote by $\mathbf{Interp}(\{(\mathbf{x}_i, y_i)\}_{i=1}^m)$[1] the set of vector fields $f \in C_1^\circ(\Omega)$ that satisfy $f(\mathbf{x}_i) = y_i$ for $i = 1, \ldots, m$. An important special case is the setting where the vectors y_i are equal to 0. In this case, the side information is the knowledge of certain equilibrium points of the vector field f.
- **Group symmetry.** For two given *linear representations*[2] $\sigma, \rho : G \to \mathbb{R}^{n \times n}$ of a finite group G, with $\sigma(g)x \in \Omega \ \forall (x, g) \in \Omega \times G$, we define $\mathbf{Sym}(G, \sigma, \rho,)$ to be the set of vector fields $f \in C_1^\circ(\Omega)$ that satisfy the symmetry condition

$$f(\sigma(g)\mathbf{x}) = \rho(g) f(\mathbf{x}) \ \forall \mathbf{x} \in \Omega, \ \forall g \in G.$$

  For example, let Ω be the unit ball in $\mathbb{R}^n$ and consider the group $F_2 = \{1, -1\}$, with scalar multiplication as the group operation. If we take σ_{F_2} and ρ_{F_2} to be two

[1] To simplify notation, we drop the dependence of the side information on the set Ω.

[2] Recall that a linear representation of a group G on the vector space $\mathbb{R}^n$ is any group homomorphism from G to the group $GL(\mathbb{R}^n)$ of invertible $n \times n$ matrices. That is, a linear representation is a map $\mu : G \to GL(\mathbb{R}^n)$ that satisfies $\mu(gg') = \mu(g)\mu(g') \ \forall g, g' \in G$.

linear representations of F_2 defined by $\sigma_{F_2}(1) = -\sigma_{F_2}(-1) = \rho_{F_2}(1) = \rho_{F_2}(-1) = I$, where I denotes the $n \times n$ identity matrix, then the set $\mathbf{Sym}(F_2, \sigma_{F_2}, \rho_{F_2},)$ (resp. $\mathbf{Sym}(F_2, \sigma_{F_2}, \sigma_{F_2},)$) is exactly the set of even (resp. odd) vector fields in $C_1^\circ(\Omega)$. As another example, consider the group S_n of all permutations of the set $\{1, \ldots, n\}$, with composition as the group operation. If we take σ_{S_n} to be the map that assigns to an element $p \in S_n$ the permutation matrix P obtained by shuffling the columns of the identity matrix according to p, and ρ_{S_n} to be the constant map that assigns the identity matrix to every $p \in S_n$, then the set $\mathbf{Sym}(S_n, \sigma_{S_n}, \rho_{S_n},)$ is the set of *symmetric* functions in $C_1^\circ(\Omega)$, i.e., functions in $C_1^\circ(\Omega)$ that are invariant under permutations of their arguments. We remark that a finite combination of side information of type **Sym** can often be written equivalently as a single side information of type **Sym**. For example, the set $\mathbf{Sym}(F_2, \sigma_{F_2}, \rho_{F_2}, \cap)\, \mathbf{Sym}(S_n, \sigma_{S_n}, \rho_{S_n},)$ of even symmetric functions in $C_1^\circ(\Omega)$ is equal to $\mathbf{Sym}(F_2 \times S_n, \sigma, \rho,)$, where $F_2 \times S_n$ is the direct product of F_2 and S_n, σ is given by $\sigma(g, g') = \sigma_{F_2}(g)\sigma_{S_n}(g')\, \forall (g, g') \in F_2 \times S_n$, and ρ is the constant map that assigns the identity matrix to every element in $F_2 \times S_n$.

- **Coordinate nonnegativity.** For given sets $P_i, N_i \subseteq \Omega$, $i = 1, \ldots, n$, we denote by $\mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$ the set of vector fields $f \in C_1^\circ(\Omega)$ that satisfy

$$f_i(\mathbf{x}) \geq 0\ \forall \mathbf{x} \in P_i, \text{ and } f_i(\mathbf{x}) \leq 0\ \forall \mathbf{x} \in N_i,\ \forall i \in \{1, \ldots, n\}.$$

  These constraints are useful when we know that certain components of the state vector are increasing or decreasing functions of time in some regions of the state space.[3]

- **Coordinate directional monotonicity.** For given sets $P_{ij}, N_{ij} \subseteq \Omega$, $i, j = 1, \ldots, n$, we denote by $\mathbf{Mon}(\{(P_{ij}, N_{ij})\}_{i,j=1}^n)$ the set of vector fields $f \in C_1^\circ(\Omega)$ that satisfy

$$\frac{\partial f_i}{\partial x_j}(\mathbf{x}) \geq 0\ \forall \mathbf{x} \in P_{ij}, \text{ and } \frac{\partial f_i}{\partial x_j}(\mathbf{x}) \leq 0\ \forall \mathbf{x} \in N_{ij},\ \forall i, j \in \{1, \ldots, n\}.$$

  See fig. 6.2 for an illustration of a simple example.

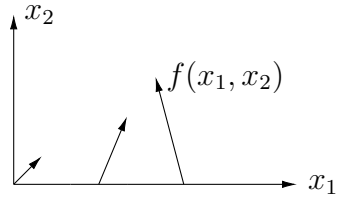


Figure 6.2: An example of the behavior of a vector field $f : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying $\mathbf{Mon}(\{(P_{ij}, N_{ij})\}_{i,j=1}^2)$ with $P_{21} = N_{11} = [0, 1] \times \{0\}$ (i.e., $\frac{\partial f_2}{\partial x_1}(x_1, 0) \geq 0$ and $\frac{\partial f_1}{\partial x_1}(x_1, 0) \leq 0\ \forall x_1 \in [0, 1]$), and with the rest of the sets P_{ij} and N_{ij} equal to the empty set.

[3] There is no loss of generality in assuming that each coordinate of the vector field is nonnegative or nonpositive on a single set since one can always reduce multiple sets to one by taking unions. The same comment applies to the side information of coordinate directional monotonicity that is defined next.

In the special case where $P_{ij} = \mathbb{R}^n$ and $N_{ij} = \emptyset$ for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$, and $P_{ii} = N_{ii} = \emptyset$ for all $i \in \{1, \ldots, n\}$, the side information is the knowledge of the following property of the vector field f which appears frequently in the literature on monotone systems [193]:

$$\forall x_{\text{init}}, \tilde{x}_{\text{init}} \in \mathbb{R}^n, \quad x_{\text{init}} \leq \tilde{x}_{\text{init}} \implies x(t, x_{\text{init}}) \leq x(t, \tilde{x}_{\text{init}}) \; \forall t \geq 0.$$

Here, the inequalities are interpreted elementwise, and the notation $x(t, x_{\text{init}})$ is used as before to denote the trajectory of the vector field f starting from the initial condition x_{init}.

- **Invariance of a set.** A set $B \subseteq \Omega$ is invariant under a vector field f if any trajectory of the dynamical system $\dot{\mathbf{x}}(t) = f(\mathbf{x}(t))$ which starts in B stays in B forever. In particular, if $B = \{x \in \mathbb{R}^n \mid h_j(\mathbf{x}) \geq 0, \; j = 1, , \ldots, m\}$ for some differentiable functions $h_j : \mathbb{R}^n \to \mathbb{R}$, then invariance of the set B under the vector field f implies the following constraints (see Figure 6.3 for an illustration):

$$\forall j \in \{1, \ldots, m\}, \; \forall x \in B, \quad [h_j(x) = 0 \implies \langle f(\mathbf{x}), \nabla h_j(\mathbf{x}) \rangle \geq 0] \,. \tag{6.7}$$

Indeed, suppose for some $\tilde{x} \in B$ and for some $j \in \{1, \ldots, m\}$, we had $h_j(\tilde{x}) = 0$ but $\langle f(\tilde{x}), \nabla h_j(\tilde{x}) \rangle = \dot{h}(\tilde{x}) < 0$, then $h(x(t, \tilde{x})) < 0$ for t small enough, implying that $x(t, \tilde{x}) \notin B$ for t small enough. It is also straightforward to verify that if the "$\geq$" in (6.7) were replaced with a "$>$", then the resulting condition would be sufficient for invariance of the set B under f. In fact, it follows from a theorem of Nagumo [145, 43] that condition (6.7) is necessary and sufficient for invariance of the set B under f if B is convex, the functions $h_1, \ldots, h_m$ are continuously-differentiable, and for every point x on the boundary of B, the vectors $\{\nabla h_j(x) \mid j \in \{1, \ldots, m\}, h_j(x) = 0\}$ are linearly independent.[4] Given sets $B_i = \{x \in \mathbb{R}^n \mid h_{ij}(\mathbf{x}) \geq 0, \; j = 1, , \ldots, m_i\}$, $i = 1, \ldots, r$, defined by differentiable functions $h_{ij} : \mathbb{R}^n \to \mathbb{R}$, we denote by $\mathbf{Inv}(\{B_i\}_{i=1}^r)$ the set of all vector fields $f \in C_1^\circ(\Omega)$ that satisfy (6.7) for $B \in \{B_1, \ldots, B_r\}$.

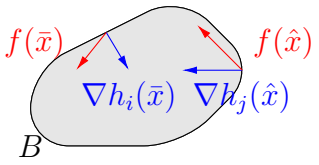


Figure 6.3: An example of the behavior of a vector field $f : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying $\mathbf{Inv}(\{B\})$, where $B := \{x \in \mathbb{R}^2 \mid h_1(x) \geq 0, \ldots, h_m(x) \geq 0\}$ is the set shaded in gray.

- **Gradient and Hamiltonian systems.** A vector field $f \in C_1^\circ(\Omega)$ is said to be a *gradient* vector field if there exists a differentiable, scalar-valued function $V : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(\mathbf{x}) = -\nabla V(\mathbf{x}) \; \forall \mathbf{x} \in \Omega. \tag{6.8}$$

Typically, the function V is interpreted as a notion of potential or energy associated with the dynamical system $\dot{x}(t) = f(x(t))$. Note that the value of the function V

[4] In an earlier draft of this paper [6], we had incorrectly claimed that condition eq. (6.7) is equivalent to invariance of the set B, while in fact additional assumptions are needed for its sufficiency.

decreases along the trajectories of this dynamical system. We denote by **Grad** the subset of $C_1^\circ(\Omega)$ consisting of gradient vector fields.

A vector field $f \in C_1^\circ(\Omega)$ over n state variables $(x_1, \ldots, x_n)$ is said to be *Hamiltonian* if n is even and there exists a differentiable scalar-valued function $H : \mathbb{R}^n \to \mathbb{R}$ such that

$$f_i(p,q) = -\frac{\partial H}{\partial q_i}(p,q), f_{\frac{n}{2}+i}(p,q) = \frac{\partial H}{\partial p_i}(p,q),\ \forall (p,q) \in \Omega,\ \forall i \in \left\{1, \ldots, \frac{n}{2}\right\},$$

where $p := (x_1, \ldots, x_{\frac{n}{2}})^T$ and $q := (\mathbf{x}_{\frac{n}{2}+1}, \ldots, \mathbf{x}_n)^T$. The states p and $\mathbf{q}$ are usually referred to as *generalized momentum* and *generalized position* respectively, following terminology from physics. Note that a Hamiltonian system conserves the quantity H along its trajectories. We denote by **Ham** the subset of $C_1^\circ(\Omega)$ consisting of Hamiltonian vector fields. For related work on learning Hamiltonian systems, see [14, 85].

6.3 Learning Polynomial Vector Fields Subject to Side Information

In this paper, we take the function class $\mathcal{F}$ in (6.6) to be the set of polynomial vector fields of a given degree d. We denote this function class by

$$\mathcal{P}_d := \{p : \mathbb{R}^n \to \mathbb{R}^n \mid p_i \text{ is a (scalar-valued) polynomial of degree } d \text{ for } i = 1, \ldots, n\}.$$

Furthermore, we assume that the set Ω over which we would like to learn the unknown dynamical system, the sets P_i, N_i in the definition of $\mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$, the sets P_{ij}, N_{ij} in the definition of $\mathbf{Mon}(\{P_{ij}, N_{ij}\}_{i,j=1}^n)$, and the sets B_i in the definition of $\mathbf{Inv}(\{B_i\}_{i=1}^r)$ are all closed semialgebraic. We recall that a *closed basic semialgebraic* set is a set of the form

$$\Lambda := \{x \in \mathbb{R}^n |\ g_i(x) \geq 0, i = 1, \ldots, m\}, \tag{6.9}$$

where $g_1, \ldots, g_m$ are (scalar-valued) polynomial functions, and that a *closed semialgebraic* set is a finite union of closed basic semialgebraic sets.

Our choice of working with polynomial functions to describe the vector field and the sets that appear in the side information definitions are motivated by two reasons. The first is that polynomial functions are expressive enough to represent or approximate a large family of functions and sets that appear in applications. The second reason, which shall be made clear shortly, is that because of some connections between real algebra and semidefinite optimization, several side information constraints that are commonly available in practice can be imposed on polynomial vector fields in a numerically tractable fashion.

With our aforementioned choices, the optimization problem in (6.6) has as decision variables the coefficients of a candidate polynomial vector field $p : \mathbb{R}^n \to \mathbb{R}^n$. When the notion of side information is restricted to the six types presented in Section 6.2,

and under the mild assumptions that Ω is full dimensional (i.e, that it contains an open set), the constraints of (6.6) are of the following two types:

(i) Affine constraints in the coefficients of p.

(ii) Constraints of the type

$$q(x) \geq 0 \ \forall x \in \Lambda, \tag{6.10}$$

where Λ is a given closed basic semialgebraic set of the form (6.9), and q is a (scalar-valued) polynomial whose coefficients depend affinely on the coefficients of the vector field p.

For example, membership to $\mathbf{Interp}(\{(x_i, y_i)\}_{i=1}^m)$, $\mathbf{Sym}(G, \sigma, \rho,)$, $\mathbf{Grad}$, or $\mathbf{Ham}$ can be enforced by affine constraints,[5] while membership to $\mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$, $\mathbf{Inv}(\{B_i\}_{i=1}^r)$, or $\mathbf{Mon}(\{(P_{ij}, N_{ij})\}_{i,j=1}^n)$ can be cast as constraints of the type (6.10). Unfortunately, imposing the latter type of constraints is NP-hard already when q is a quartic polynomial and $\Lambda = \mathbb{R}^n$, or when q is quadratic and Λ is a polytope (see, e.g., [144]).

An idea pioneered to a large extent by Lasserre [122] and Parrilo [157] has been to write algebraic sufficient conditions for (6.10) based on the concept of sum of squares polynomials. We say that a polynomial h is a *sum of squares* (sos) if it can be written as $h = \sum_i q_i^2$ for some polynomials q_i. Observe that if we succeed in finding sos polynomials $\sigma_0, \sigma_1, \ldots, \sigma_m$ such that the polynomial identity

$$q(x) = \sigma_0(x) + \sigma_1(x)g_1(x) + \ldots + \sigma_m(x)g_m(x) \tag{6.11}$$

holds (for all $x \in \mathbb{R}^n$), then, clearly, the constraint in (6.10) must be satisfied. When the degrees of the sos polynomials σ_i are bounded above by an integer r, we refer to the identity in (6.11) as the *degree-r sos certificate* of the constraint in (6.10). Conversely, the following celebrated result in algebraic geometry [172] states that if $g_1, \ldots, g_m$ satisfy the so-called "Archimedean property" (a condition slightly stronger than compactness of the set Λ), then positivity of q on Λ guarantees existence of a degree-r sos certificate for some integer r large enough.

Theorem 6.3.1 (Putinar's Positivstellensatz [172])**.** *Let*

$$\Lambda = \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$$

and assume that the collection of polynomials $\{g_1, \ldots, g_m\}$ satisfies the Archimedean property, i.e., there exists a positive scalar R such that

$$R^2 - \sum_{i=1}^n x_i^2 = s_0(x) + s_1(x)g_1(x) + \ldots + s_m(x)g_m(x),$$

[5] To see why membership of a polynomial vector field p to $\mathbf{Grad}$ can be enforced by affine constraints (a similar argument works for membership to $\mathbf{Ham}$), observe that if there exists a continuously-differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ such that $p(x) = -\nabla V(x)$ for all x in a full-dimensional set Ω, then the function V is necessarily a polynomial of degree equal to the degree of p plus one. Furthermore, equality between two polynomial functions over a full-dimensional set can be enforced by equating their coefficients.

where $s_0, \ldots, s_m$ are sos polynomials.[6] *For any polynomial q, if $q(x) > 0 \; \forall \; x \in \Lambda$, then*

$$q(x) = \sigma_0(x) + \sigma_1(x)g_1(x) + \ldots + \sigma_m(x)g_m(x),$$

for some sos polynomials $\sigma_0, \ldots, \sigma_m$.

The computational appeal of the sum of squares approach stems from the fact that the search for sos polynomials $\sigma_0, \sigma_1, \ldots, \sigma_m$ of a given degree that verify the polynomial identity in (6.11) can be automated via *semidefinite programming* (SDP)[7]. This is true even when some coefficients of the polynomial q are left as decision variables. This claim is a straightforward consequence of the following well-known fact (see, e.g., [156]): A polynomial h of degree $2d$ is a sum of squares if and only if there exists a symmetric matrix Q which is positive semidefinite and verifies the identity

$$h(x) = z(x)^T Q z(x), \tag{6.12}$$

where $z(x)$ denotes the vector of all monomials in x of degree less than or equal to d. Identity (6.12) can be written in an equivalent manner as a system of $\binom{n+d}{d}$ linear equations involving the entries of the matrix Q and the coefficients of the polynomial h. These equations come from equating the coefficients of the polynomials appearing on the left and right hand sides of (6.12). The problem of finding a positive semidefintie matrix Q whose entries satisfy these linear equations is a semidefinite program. For implementation purposes, there exist modeling languages, such as YALMIP [131], SOSTOOLS [166], or SumOfSquares.jl [208], that accept sos constraints on polynomials directly and do the conversion to a semidefinite program in the background. See e.g. [126, 46, 90] for more background on sum of squares techniques.

To end up with a semidefinite programming formulation of problem (6.6), we also need to take the loss function ℓ that appears in the objective function to be semidefinite representable (i.e., we need its epigraph to be the projection of the feasible set of a semidefinite program). Luckily, many common loss functions in machine learning are semidefinite representable. Examples of such loss functions include (i) any ℓ_p norm for a rational number $p \geq 1$, or for $p = \infty$, (ii) any convex piece-wise linear function, (iii) any *sos-convex* polynomial (see e.g. [93] for a definition), and (iv) any positive integer power of the previous three function classes.

6.4 Illustrative Experiments

In this section, we present numerical experiments from four application domains to illustrate our methodology. The first three applications are learning experiments and

[6] If Λ is known to be contained in a ball of radius R, one can add the redundant constraint $R^2 - \sum_{i=1}^n x_i^2 \geq 0$ to the description of Λ, and then the Archimedean property will be automatically satisfied.

[7] Semidefinite programming is the problem of minimizing a linear function of a symmetric matrix over the intersection of the cone of positive semidefinite matrices with an affine subspace. Semidefinite programs can be solved to arbitrary accuracy in polynomial time; see [202] for a survey of the theory and applications of this subject.

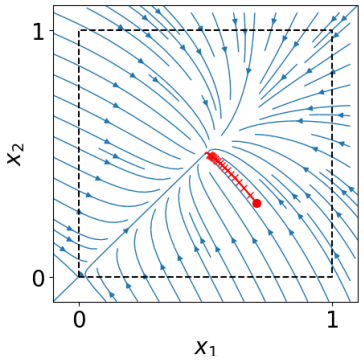


Figure 6.4: Streamplot of the vector field in (6.13). We consider this vector field to be the ground truth and unknown to us. We would like to learn it over $[0,1]^2$ from noisy snapshots of a single trajectory starting from $(0.7, 0.3)^T$ (plotted inred).

the last one involves an optimal control component. In all of our experiments, we use the SDP-based approach explained in Section 6.3 to tackle problem (6.6) and take our loss function $\ell(\cdot,\cdot)$ in (6.6) to be $\ell(u,v) = \|u-v\|_2^2 \quad \forall u, v \in \mathbb{R}^n$. The added value of side information for learning dynamical systems from data will be demonstrated in these experiments.

6.4.1 Diffusion of a contagious disease

The following dynamical system has appeared in the epidemiology literature (see, e.g., [18]) as a model for the spread of a sexually transmitted disease in a heterosexual population:

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)), \text{where } \mathbf{x}(t) \in \mathbb{R}^2 \text{ and } f(\mathbf{x}) = \begin{pmatrix} -a_1 x_1 + b_1(1-x_1)x_2 \\ -a_2 x_2 + b_2(1-x_2)x_1 \end{pmatrix}. \tag{6.13}$$

Here, the quantity $x_1(t)$ (resp. $x_2(t)$) represents the fraction of infected males (resp. females) in the population. The parameter a_i (resp. b_i) denotes the recovery rate (resp. the infection rate) in the male population when $i = 1$, and in the female population when $i = 2$. We take

$$(a_1, b_1, a_2, b_2) = (0.05, 0.1, 0.05, 0.1) \tag{6.14}$$

and plot the resulting vector field f in fig. 6.4. We suppose that this vector field is *unknown* to us, and our goal is to learn it over $\Omega := [0,1]^2$ from a few noisy snapshots of a single trajectory. More specifically, we have access to the training data set

$$\mathcal{D} := \left\{ \left(\mathbf{x}(t_i, x_{\text{init}}), f(\mathbf{x}(t_i, x_{\text{init}})) + 10^{-4} \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \end{pmatrix} \right) \right\}_{i=1}^{20}, \tag{6.15}$$

where $\mathbf{x}(t, x_{\text{init}})$ is the trajectory of the system (6.13) starting from the initial condition $\mathbf{x}_{\text{init}} = (0.7, 0.3)^T$, the scalars $t_i := i$ represent a uniform subdivision of the time interval $[0, 20]$, and the scalars $\varepsilon_{1,1}, \varepsilon_{1,2}, \ldots, \varepsilon_{20,1}, \varepsilon_{20,2}$ are independently sampled from the standard normal distribution.

Following our approach in Section 6.3, we parameterize our candidate vector field $p : \mathbb{R}^2 \to \mathbb{R}^2$ as a polynomial function. We choose the degree of this polynomial to be $d = 3$. The degree d is taken to be larger than 2 because we do not want to assume knowledge of the degree of the true vector field in (6.13). This makes the task of learning more difficult; see the end of this subsection where we also learn a vector field of degree 2 for comparison.

In absence of any side information, one could solve the least-squares problem

$$\min_{p \in \mathcal{P}_d} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} \|p(\mathbf{x}_i) - \mathbf{y}_i\|_2^2 \tag{6.16}$$

to find a polynomial of degree d that best agrees with the training data. For this experiment only, and for educational purposes, we include a template code using the library SumOfSquares.jl [208] of the Julia programming language and demonstrate how the code changes as we impose side information constraints. We initiate our template with the following code that solves optimization problem eq. (6.16).

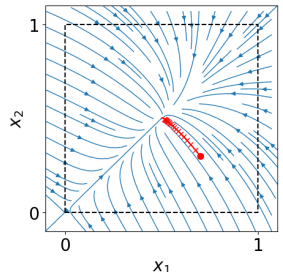
```
# Input: vectors X1, X2, Y1, Y2 ∈ ℝ^20 representing the training set in (6.15),
#        where X1 = {x1(ti, xinit)}_{i=1}^20, X2 = {x2(ti, xinit)}_{i=1}^20,
#        Y1 = {f1(x(ti, xinit)) + 10^-4 εi,1}_{i=1}^20, and Y2 = {f2(x(ti, xinit)) + 10^-4 εi,2}_{i=1}^20

model = SOSModel(solver)                          # solver could be any SDP solver,
                                                  # e.g., Mosek [22], SDPT3 [199], CSDP [48]
@polyvar x1 x2                                    # Define state variables x1, x2
d = 3                                             # Construct vector of monomials
z = monomials([x1, x2], 0:d)                      # in (x1, x2) up to degree d
@variable(model, p1, Poly(z))                     # Declare a polynomial vector field
@variable(model, p2, Poly(z))                     # whose coefficients are decision variables
error_vec = [p[1](X1, X2) - Y1;                   # Vector of individual terms appearing
             p[2](X1, X2) - Y2]                   # in the objective of eq. (6.16)
@objective model Min error_vec' * error_vec
# Side information constraints go here
# ...
optimize!(model)                                  # Solve the optimization problem
```

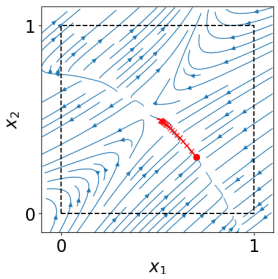
Julia template code for learning dynamical systems with side information.

The solution to problem (6.16) returned by the solver MOSEK [22] is plotted in fig. 6.5b. Observe that while the learned vector field replicates the behavior of the true vector field f on the observed trajectory, it differs significantly from f on the rest of the unit square. To remedy this problem, we leverage the following list of side information that is available from the context without knowing the exact structure of f.

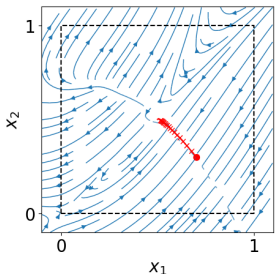
- **Equilibrium point at the origin (Interp).** Naturally, if no male or female is infected, there would be no contagion and the number of infected individuals will



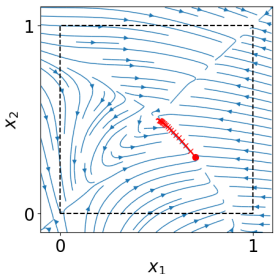
(a) The true vector field (unknown to us)



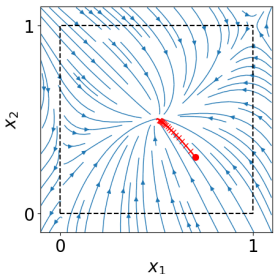
(b) No side information



(c) **Interp**



(d) **Interp** $\cap$ **Inv**



(e) **Interp** $\cap$ **Inv** $\cap$ **Mon**

Figure 6.5: Streamplot of the vector field in (6.13) (fig. 6.5a) along with streamplots of polynomial vector fields of degree 3 that are optimal to (6.16) with different side information constraints appended to it (figs. 6.5b to 6.5e).

remain at zero. This side information corresponds to our vector field p having an equilibrium point at the origin, i.e., $p(0,0) = 0$. Note from figs. 6.5a and 6.5b that the true vector field f in (6.13) satisfies this constraint, but the vector field learned by solving the least-squares problem in (6.16) does not. We can impose this linear constraint by simply adding the following lines of code to our template:

```
@constraint model p1(0,0) == 0
@constraint model p2(0,0) == 0
```

The vector field resulting from solving this new problem is plotted in fig. 6.5c.

- **Invariance of the box $[0,1]^2$ (Inv).** The state variables (x_1, x_2) of the dynamics in (6.13) represent fractions of infected individuals and as such, the vector $\mathbf{x}(t)$ should be contained in the box $[0,1]^2$ at all times $t \geq 0$. Note that this property is violated by the vector fields learned in figs. 6.5b and 6.5c. Mathematically, the invariance of the unit box is equivalent to the four (univariate) polynomial nonnegativity constraints

$$p_2(x_1, 0) \geq 0, p_2(x_1, 1) \leq 0 \; \forall x_1 \in [0,1],$$

$$p_1(0, x_2) \geq 0, p_1(1, x_2) \leq 0 \; \forall x_2 \in [0,1].$$

These constraints imply that the vector field points inwards on the four edges of the unit box. We replace each one of these four constraints with the corresponding degree-2 sos certificate of the type in (6.11). For instance, we replace the constraint

$$p_2(x_1, 0) \geq 0 \; \forall x_1 \in [0, 1]$$

with linear and semidefinite constraints obtained from equating the coefficients of the two sides of the polynomial identity

$$p_2(x_1, 0) = x_1 s_0(x_1) + (1 - x_1) s_1(x_1), \tag{6.17}$$

and requiring that the newly-introduced (univariate) polynomials s_0 and s_1 be quadratic and sos. Obviously, the algebraic identity (6.17) is sufficient for nonnegativity of $p_2(x_1, 0)$ over $[0, 1]$; in this case, it also happens to be necessary [134]. The code in Julia for imposing the degree-2 sos certificate in (6.17) is as follows:

```
@variable(model, s1, Poly([1, x1, x1^2]))      # Declare decision polynomial s1
@variable(model, s2, Poly([1, x1, x1^2]))      # Declare decision polynomial s2
@constraint(model, s1, in SOSCone())           # Enforce s1 to be sos
@constraint(model, s2, in SOSCone())           # Enforce s2 to be sos

polynomial_identity =                          # Enfroce polynomial
  p2((x1, x2) ⇒ (0, x2)) - x1*s1-(1-x1)*s2     # identity in eq. (6.17)

@constraint(model, coefficients(polynomial_identity).== 0)
```

The output of the semidefinite program which imposes the invariance of the unit box and the equilibrium point at the origin is plotted in fig. 6.5d.

- **Coordinate directional monotonicity (Mon).** Naturally, one would expect that if the fraction of infected males rises in the population, the rate of infection of females should increase. Mathematically, this observation is equivalent to the constraint that

$$\frac{\partial p_2}{\partial x_1}(\mathbf{x}) \geq 0 \; \forall \mathbf{x} \in [0, 1]^2.$$

Similarly, swapping the roles played by males and females leads to the constraint

$$\frac{\partial p_1}{\partial x_2}(\mathbf{x}) \geq 0 \; \forall \mathbf{x} \in [0, 1]^2.$$

Note that this property is violated by the vector fields learned in figs. 6.5b to 6.5d. Just as in the previous bullet point, we replace each of the above two nonnegativity constraints with its corresponding degree-2 sos certificate. To do this, we represent the closed basic semialgebraic set $[0, 1]^2$ with the polynomial inequalities

$$x_1 \geq 0, x_2 \geq 0, 1 - x_1 \geq 0, 1 - x_2 \geq 0.$$

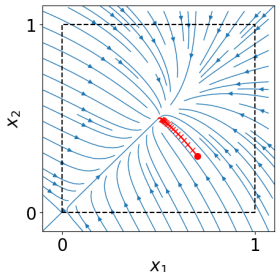
The Julia code for imposing this side information is similar to the one of the previous bullet point and therefore omitted. fig. 6.5e demonstrates the vector field learned by our semidefinite program when all side information constraints discussed thus far are imposed.

Note from figs. 6.5b to 6.5e that as we add more side information, the learned vector field respects more and more properties of the true vector field f. In particular, the learned vector field in fig. 6.5e is quite similar qualitatively to the true vector field in fig. 6.5a even though only noisy snapshots of a single trajectory are used for learning.

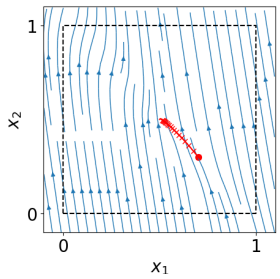
It is interesting to observe what would happen if we try to learn a degree-2 vector field from the same training set using the list of side information discussed in this subsection. The outcome of this experiment is plotted in fig. 6.6. Note that with the equilibrium-at-the-origin side information, the behavior of the learned vector field of degree 2 is already quite close to that of the true vector field. fig. 6.6e shows that when we impose all side information, the learned vector field is almost indistinguishable from the true vector field (even though, once again, only noisy snapshots of a single trajectory are used for learning). The vector field plotted in fig. 6.6e is given by

$$p_{\mathbf{Interp}\cap\mathbf{Inv}\cap\mathbf{Mon},\deg 2}(x_1,x_2)=\begin{pmatrix}0.038x_1^2-0.100x_1x_2-0.009x_2^2-0.084x_1+0.119x_2\\ -0.101x_1x_2+0.003x_2^2+0.101x_1-0.052x_2\end{pmatrix},$$

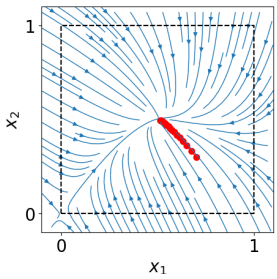
which is indeed very close to the vector field in (6.13).



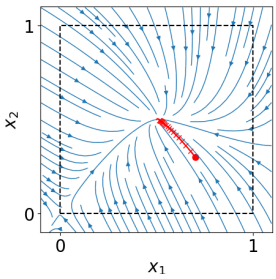
(a) The true vector field (unknown to us)



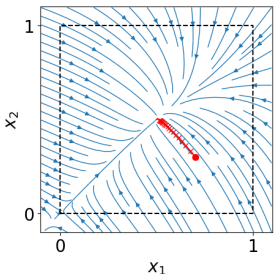
(b) No side information



(c) **Interp**



(d) **Interp** ∩ **Inv**



(e) **Interp** ∩ **Inv** ∩ **Mon**

Figure 6.6: Streamplot of the vector field in (6.13) (fig. 6.5a) along with streamplots of polynomial vector fields of degree 2 that are optimal to (6.16) with different side information constraints appended to it (figs. 6.6b to 6.6e).

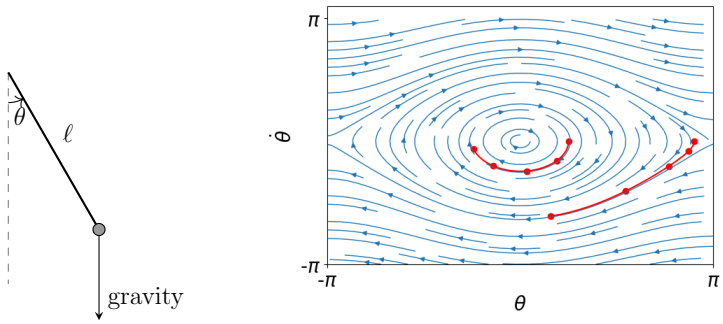


Figure 6.7: The simple pendulum and the streamplot of its vector field. We would like to learn this vector field over $[-\pi, \pi]^2$ from 10 noisy snapshots coming from two trajectories.

6.4.2 Dynamics of the simple pendulum

In this subsection, we consider the dynamics of the simple pendulum, i.e., a mass m hanging from a massless rod of length ℓ (see fig. 6.7). The state variables of this system are given by $x = (\theta, \dot{\theta})$, where θ is the angle that the rod makes with the vertical axis and $\dot{\theta}$ is the time derivative of this angle. By convention, the angle $\theta \in [-\pi, \pi]$ is positive when the mass is to the right of the vertical axis, and negative otherwise. By applying Newton's second law of motion, the equation

$$\ddot{\theta}(t) = -\frac{g}{\ell} \sin(\theta(t))$$

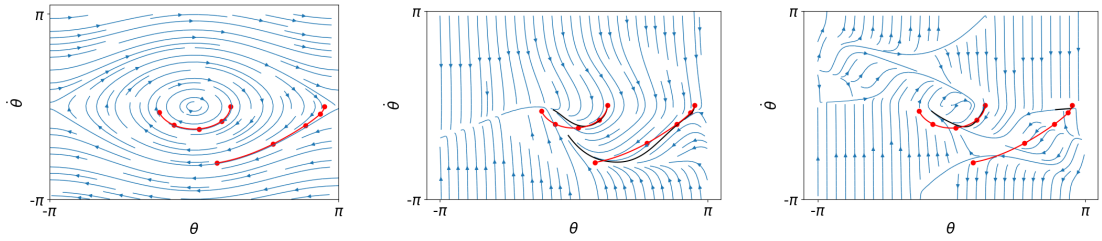
for the dynamics of the pendulum can be derived, where g here is the acceleration due to gravity. This is a one-dimensional second-order system that we convert to a first-order system as follows:

$$\dot{x}(t) = f(x(t)) \quad \text{where} \quad x(t) := \begin{pmatrix} \theta(t) \\ \dot{\theta}(t) \end{pmatrix} \text{ and } f(\theta, \dot{\theta}) := \begin{pmatrix} \dot{\theta} \\ -\frac{g}{\ell} \sin \theta \end{pmatrix}. \tag{6.18}$$

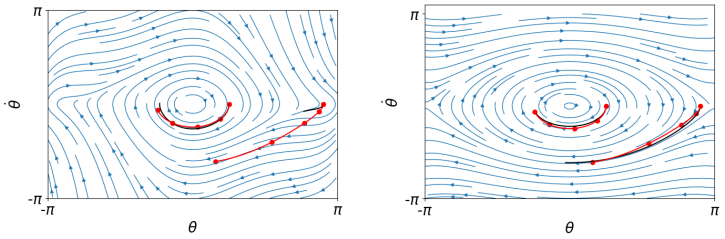
We take the vector field in (6.18) with $g = \ell = 1$ to be the ground truth. We observe from this vector field a noisy version of two trajectories $x(t, x^1_{\text{init}})$ and $x(t, x^2_{\text{init}})$ sampled at times $t_i = \frac{3i}{5}$, where $i \in \{0, \ldots, 4\}$, with $x^1_{\text{init}} = (\frac{\pi}{4}, 0)^T$ and $x^2_{\text{init}} = (\frac{9\pi}{10}, 0)^T$ (see fig. 6.7). More precisely, we assume that our training set (with a slightly different representation of its elements) is given by

$$\mathcal{D} := \bigcup_{k=1}^{2} \left\{ \begin{pmatrix} \theta(t_i, x^k_{\text{init}}) \\ \dot{\theta}(t_i, x^k_{\text{init}}) \\ \ddot{\theta}(t_i, x^k_{\text{init}}) \end{pmatrix} + 10^{-2} \varepsilon_{i,k} \right\}_{i=0}^{4}, \tag{6.19}$$

where the $\varepsilon_{i,k}$ (for $k = 1, 2$ and $i = 0, \ldots, 4$) are independent 3×1 standard normal vectors.



(a) The true vector field (unkown to us) (b) No side information (c) **Sym**



(d) **Sym** ∩ **Pos** (e) **Sym** ∩ **Pos** ∩ **Ham**

Figure 6.8: Streamplot of the vector field in (6.18) (fig. 6.8a) along with streamplots of polynomial vector fields of degree 5 that best agree with the data (in the least-squares sense) and obey an increasing number of side information constraints (figs. 6.8b to 6.8e). In each case, the trajectories of the learned vector field starting from the same two initial conditions as the trajectories observed in the training set are plotted in black.

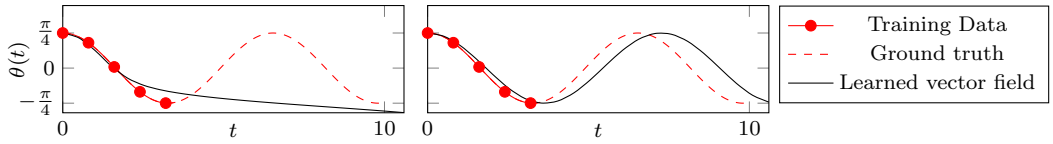


Figure 6.9: Comparison of the trajectory of the simple pendulum in (6.18) starting from $(\frac{\pi}{4}, 0)^T$ (dotted) with the trajectory from the same initial condition of the polynomial vector field of degree 5 that best agrees with the data (in the least-squares solution) in the absence of side information (left), and subject to side information constraints **Sym** ∩ **Pos** ∩ **Ham** (right).

We are interested in learning the vector field f over the set $\Omega := [-\pi, \pi]^2$ from the training data in (6.19). We parameterize our candidate vector field $p : \mathbb{R}^2 \to \mathbb{R}^2$ as a degree-5 polynomial. Note that $p_1(\theta, \dot{\theta}) = \dot{\theta}$, just from the meaning of our state variables. The only unknown is therefore the polynomial $p_2(\theta, \dot{\theta})$.

In absence of side information, one can solve a least-squares problem that finds a polynomial of degree 5 that best agrees with the training data. As it can be seen in fig. 6.8b, the resulting vector field is very far from the true vector field and is unable to even replicate the observed trajectories when started from the same two initial

conditions. To learn a better model, we describe a list of side information which could be derived from contextual knowledge without knowing the true vector field f.

- **Sign symmetry (Sym).** The pendulum obviously behaves symmetrically with respect to the vertical axis (plotted with a dotted line in fig. 6.7). We therefore require our candidate vector field p to satisfy the same symmetry condition

$$p(-\theta, -\dot{\theta}) = -p(\theta, \dot{\theta}) \quad \forall (\theta, \dot{\theta}) \in \Omega.$$

  Note that this is an affine constraint in the coefficients of the polynomial p, and that the true vector field f in (6.18) satisfies this constraint.

- **Coordinate nonnegativity (Pos).** We know that the force of gravity pulls the pendulum's mass down and pushes the angle θ towards 0. This means that the angular velocity $\dot{\theta}$ decreases when θ is positive and increases when θ is negative. Mathematically, we must have

$$p_2(\theta, \dot{\theta}) \leq 0 \; \forall (\theta, \dot{\theta}) \in [0, \pi] \times [-\pi, \pi] \text{ and } p_2(\theta, \dot{\theta}) \geq 0 \; \forall (\theta, \dot{\theta}) \in [-\pi, 0] \times [-\pi, \pi].$$

  We replace each one of these constraints with their corresponding degree-4 sos certificate (see the definition following equation (6.11)). (Note that, because of the previous symmetry side information, we actually only need to impose one of these two constraints.)

- **Hamiltonian (Ham).** In the simple pendulum model, there is no dissipation of energy (through friction for example), so the total energy

$$E(\theta, \dot{\theta}) = \frac{1}{2} m l^2 \dot{\theta}^2 + mgl(1 - \cos(\theta)) \tag{6.20}$$

  is conserved. The two terms appearing in this equation can be interpreted physically as the kinetic and the potential energy of the system. Furthermore, the total energy E satisfies

$$\dot{\theta}(t) = \frac{1}{m\ell^2} \frac{\partial E}{\partial \dot{\theta}}(\theta(t), \dot{\theta}(t)), \text{ and } \ddot{\theta}(t) = -\frac{1}{m\ell^2} \frac{\partial E}{\partial \theta}(\theta(t), \dot{\theta}(t)).$$

  The simple pendulum system is therefore a Hamiltonian system, with the associated Hamiltonian function $\frac{E}{m\ell^2}$. Note that neither the vector field in (6.18) describing the dynamics of the simple pendulum nor the associated Hamiltonian are polynomial functions. In our learning procedure, we use only the fact that the system is Hamiltonian, i.e., that there exists a function H such that

$$p_1(\theta, \dot{\theta}) = \frac{\partial H}{\partial \dot{\theta}}(\theta, \dot{\theta}), \text{ and } p_2(\theta, \dot{\theta}) = -\frac{\partial H}{\partial \theta}(\theta, \dot{\theta}), \tag{6.21}$$

but not the exact form of this Hamiltonian. Since we are parameterizing the candidate vector field p as a degree-5 polynomial, the function H must be a (scalar-valued) polynomial of degree 6. The Hamiltonian structure can thus be imposed by adding affine constraints on the coefficients of p, or by directly learning H and obtaining p from (6.21).

Observe from fig. 6.8 that as more side information is added, the behavior of the learned vector field gets closer to the truth. In particular, the solution returned by our semidefinite program in fig. 6.8e is almost identical to the true dynamics in fig. 6.7 even though it is obtained only from 10 noisy samples on two trajectories. fig. 6.9 shows that even if we start from an initial condition from which we have made partial trajectory observations, using side information can lead to better predictions for the future of the trajectory.

6.4.3 Growth of cancerous tumor cells

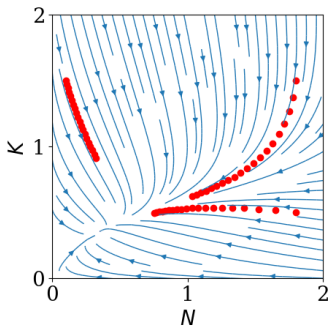


Figure 6.10: Streamplot of the vector field in eq. (6.23) describing the time evolution of the volume N of a cancerous tumor and the host's carrying capacity K (in *cubic centimeters*). We consider this vector field to be the ground truth and unkown to us. We try to learn it over $[0, 2]^2$ from noisy measurements of three trajectories (plotted in red).

In this subsection, we consider a model governing the time evolution of the volume N of a cancerous tumor inside a human body [180]. Cancerous tumors depend for their development on availability of the so-called *Endothelial* cells, the supply of which is characterized by a quantity called the *carrying capacity* K. Intuitively, K, which has the same unit as N, is proportional to the physical and energetic resources available for cell growth.

Two common modeling assumptions in cancer cell biology are that (i) the growth rate of the tumor decreases as the tumor grows, and (ii) that the volume of the tumor increases (resp. decreases) if it is below (resp. above) the carrying capacity. We follow the dynamics proposed in [180], which in contrast to prior works in that literature, also models the time evolution of the carrying capacity. The dynamics reads

$$\begin{pmatrix} \dot{N}(t) \\ \dot{K}(t) \end{pmatrix} = f(N(t), K(t)), \tag{6.22}$$

where $f : \mathbb{R}^2 \to \mathbb{R}^2$ is given by

$$f_1(N, K) := \frac{\mu N}{\nu}\left[1 - \left(\frac{N}{K}\right)^{\nu}\right], \quad f_2(N, K) := \omega N - \gamma N^{\frac{2}{3}} K. \tag{6.23}$$

Here, μ, ν, γ and ω are positive parameters. The dynamics of N is motivated by the so-called generalized logistic differential equation; the term ωN models the influence of the tumor on the Endothelial cells via short-range stimulators, and the term $-\gamma N^{\frac{2}{3}} K$ models this same influence via long-range inhibitors. See [180] for more details.

Having an accurate model of the growth dyanmics of cancerous tumors is crucial for the follow-up task of designing treatment plans, e.g., via radiation therapy. We hope that in the future leveraging side information can lead to learning more accurate models directly from patient data (as opposed to postulating an exact functional form such as (6.23)). For the moment however, we take (6.23) with the following parameters to be the ground truth:

$$(\mu, \nu, \gamma, \omega) = \frac{1}{10}(1, 5, 1, 2).$$

See fig. 6.10 for a streamplot of the corresponding vector field.

We consider the task of learning the vector field f over the compact set $\Omega := [0, 2]^2$ from noisy snapshots of three trajectories. Each trajectory was started from a random initial conditions $x^k_{\text{init}} := (N^k_{\text{init}}, K^k_{\text{init}})$ (with $k = 1, 2, 3$) inside Ω and sampled at times $t_i = i/20$, with $i = 0, \ldots, 19$ (see fig. 6.10). More precisely, we have access to the following training data:

$$\mathcal{D} := \left\{\left((N(t_i, x^k_{\text{init}}), K(t_i, x^k_{\text{init}})), (\dot{N}(t_i, x^k_{\text{init}}) + 10^{-4}\varepsilon_{i,k,1}, \dot{K}(t_i, x^k_{\text{init}}) + 10^{-4}\varepsilon_{i,k,2})\right)\right\}_{0 \le i < 20, 1 \le k \le 3}, \tag{6.24}$$

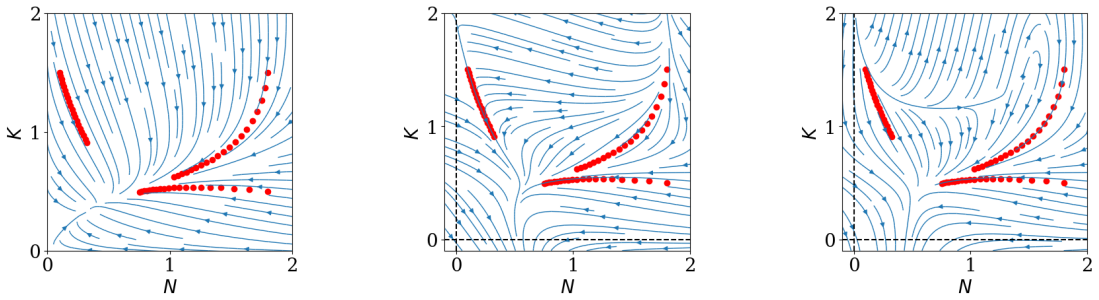
where the $\varepsilon_{i,k,l}$ (for $i = 0, \ldots, 19$, $k = 1, \ldots, 3$, and $l = 1, 2$) are independent standard normal variables. We parameterize our candidate vector field p as a degree-5 polynomial. Without any side infromation, fitting this candidate vector field to the data in eq. (6.24) via a least-squares problem leads to the vector field plotted in fig. 6.11b. Once again, the vector field obtained in this way is very far from the true vector field.

To do a better job at learning, we impose the side information constraints listed below that come from expert knowledge in the tumor growth literature (see, e.g., [180, 192, 92]):

- **A mix between coordinate nonnegativity and coordinate directional monotonicity (Pos - Mon).** As stated in [180], "one of the few near-universal observations about solid tumors is that almost all decelerate, i.e., reduce their specific growth rate $\frac{\dot{N}}{N}$, as they grow larger."

  Based on this contextual knowledge, our candidate vector field p should satisfy

$$\frac{1}{N}\frac{\partial p_1}{\partial N}(N, K) - \frac{1}{N^2} p_1(N, K) \le 0 \quad \forall N \in (0, 2], \ \forall K \in [0, 2].$$



(a) The true vector field (unknown to us)

(b) No side information

(c) **Pos - Mon**

(d) **Pos - Mon ∩ Inv**

(e) **Pos - Mon ∩ Inv ∩ Pos**

(f) **Pos - Mon ∩ Inv ∩ Pos ∩ Interp**

Figure 6.11: Streamplot of the vector field in (6.23) (fig. 6.11a) along with streamplots of polynomial vector fields of degree 5 that best agree with the data (in the least-squares sense) and obey an increasing number of side information constraints (figs. 6.11b to 6.11f).

Since the state variable N is nonnegative at all times, we can clear denominators to obtain the constraint

$$N\frac{\partial p_1}{\partial N}(N,K) - p_1(N,K) \leq 0 \; \forall (N,K) \in [0,2]^2.$$

This is a polynomial nonnegativity constraint over a closed basic semialgebraic set.

- **Invariance of the nonnegative orthant (Inv).** The state variables N and K quantify volumes, and as such, should be nonnegative at all times. This corresponds to the nonnegativity constraints

$$p_2(N,0) \geq 0 \; \forall N \in [0,2], \quad p_1(0,K) \geq 0 \; \forall K \in [0,2].$$

- **Coordinate nonnegativity (Pos).** As mentioned before, the rate of change $\dot{N}$ in the tumor volume is nonnegative when the carrying capacity K exceeds N, and nonpositive otherwise [180]. Mathematically, we must have

$$p_1(N,K) \geq 0 \quad \forall N \in [0,2], \; \forall K \in [N,2],$$

$$p_1(N, K) \leq 0 \quad \forall N \in [0, 2],\ \forall K \in [0, N].$$

- **Equilibrium point at the origin (Interp).** The tumor does not grow if the volume of cancerous cells and the carrying capacity are both zero. This corresponds to the constraint $p(0, 0) = 0$.

We observe from fig. 6.11 that as more side information is added, the behavior of the learned vector field gets closer and closer to the ground truth. In particular, the solution returned by our semidefinite program in fig. 6.11f is very close to the true dynamics in fig. 6.10.

6.4.4 Following learning with optimal control

In this subsection, we revisit the contagion dynamics (6.13) and study the effect of side information on policy decisions to contain an outbreak. Suppose that by an initial screening of a random subset of the population, it is estimated that a fraction 0.5 (resp. 0.4) of males (resp. females) are infected with the disease. We would like to contain the outbreak by performing daily widespread testing of the population. We introduce two control decision variables u_1 and u_2, representing respectively the fraction of the population of males and females that are tested per unit of time. We suppose that testing slows down the spread of the disease (due e.g. to appropriate action that can be taken on the positive cases) as follows:

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t)) - \begin{pmatrix} u_1 x_1 \\ u_2 x_2 \end{pmatrix}, \tag{6.25}$$

where $f(x(t))$ is the unknown dynamics of the spread of the disease in the absence of any control.

We suppose that the monetary cost of testing a fraction u_1 of males and u_2 of females is given by $\alpha(u_1 + u_2)$ for some known positive scalar α. Our goal is to minimize the sum

$$c(u_1, u_2) := x_1(T, \hat{x}_{\text{init}}) + x_2(T, \hat{x}_{\text{init}}) + \alpha(u_1 + u_2), \tag{6.26}$$

of the total number $x_1(T, \hat{x}_{\text{init}}) + x_2(T, \hat{x}_{\text{init}})$ of infected individuals at the end of a desired time period T, and the monetary cost $\alpha(u_1 + u_2)$ of our control law. Here, $x_1(t, \hat{x}_{\text{init}})$ and $x_2(t, \hat{x}_{\text{init}})$ evolve according to (6.25) when started from the point $\hat{x}_{\text{init}} = (0.5, 0.4)^T$. In our experiments, we take $T = 20$, $\alpha = 0.4$, and f to be the vector field in (6.13) with parameters in (6.14).

Given access to the vector field f, one could simply design an optimal control law (u_1^*, u_2^*) that minimizes the cost function $c(u_1, u_2)$ in (6.26) by gridding the control space $[0, 1]^2$, and computing $c(u_1, u_2)$ for every point of the grid. Indeed, for a given (u_1, u_2), $c(u_1, u_2)$ can be computed by simulating the dynamics in (6.25) from the initial condition $\hat{x}_{\text{init}}$. The optimal control law obtained by following this strategy is depicted in Figure 6.12 together with the graph of the function $c(u_1, u_2)$.

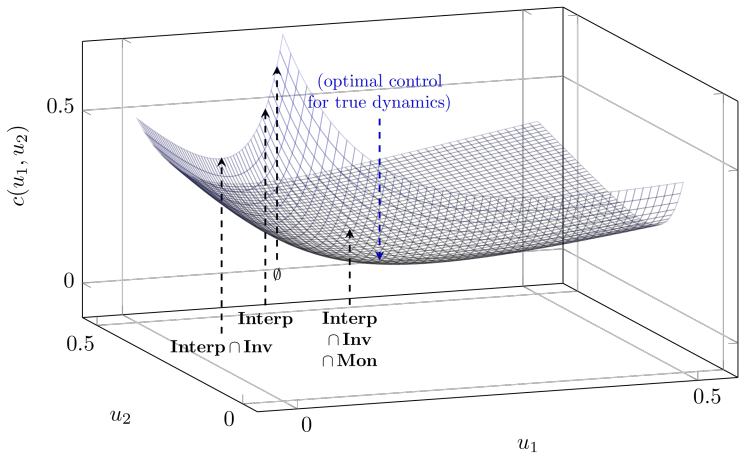


Figure 6.12: The graph of the function $c(u_1, u_2)$ in (6.26) with $T = 20$, $\alpha = 0.4$, and f as in (6.13) with parameters in (6.14). The minimizer of the function $c(u_1, u_2)$, which corresponds to the optimal control law, is indicated with a blue arrow. The control laws that are optimal for dynamics learned from a single trajectory of f with different side information constraints are indicated with black arrows.

We now consider the same setup as Section 6.4.1, where we do not know the vector field f, but have observed 20 noisy samples of a single trajectory of it starting from $(0.7, 0.3)^T$ (c.f. (6.15)). We design our control laws instead for the four polynomial vector fields of degree 3 that were learned in Section 6.4.1 with no side information, with **Interp**, with **Interp** $\cap$ **Inv**, and with **Interp** $\cap$ **Inv** $\cap$ **Mon**. This is done by following the procedure described in the previous paragraph, but using the learned vector field instead of f. The corresponding four control laws are depicted in Figure 6.12 with black arrows. We emphasize that while the control laws are computed from the learned vector fields, their associated costs in Figure 6.12 are computed by applying them to the true vector field. It is interesting to observe that adding side information constraints during the learning phase leads to the design of better control laws.

From Table 6.1, we see that if we had access to the true vector field, an optimal control law would lead to the eradication of the disease by time T. The first four rows of this table demonstrate the fraction of infected males and females at time T when control laws that are optimal for dynamics learned with different side information constraints are applied to the true vector field. It is interesting to note that with no side information, a large fraction of the population remains infected, whereas control laws computed with three side information constraints are able to eradicate the disease almost completely.

6.5 Approximation Results

In this section, we present some density results for polynomial vector fields that obey side information. This provides some theoretical justification for our choice of parameterizing our candidate vector fields as polynomial functions.

Side information	$x_1(T, \hat{x}_{\text{init}})$	$x_2(T, \hat{x}_{\text{init}})$
None	0.45	0.40
Interp	0.41	0.29
Interp $\cap$ **Inv**	0.31	0.12
Interp $\cap$ **Inv** $\cap$ **Mon**	0.01	0.01
True vector field	0.00	0.00

Table 6.1: The first four rows indicate the fraction of infected males and females at the end of the period T when a control law, optimal for dynamics learned from a single trajectory with different side information constraints, is applied to the true vector field. The last row indicates the fraction of infected males and females at time T resulting from applying the optimal control law computed with access to the true dynamics.

More precisely, we are interested in the following question: Given a continuously-differentiable vector field f satisfying a list of side information constraints from Section 6.2, is there a polynomial vector field that is "close" to f and satisfies the same list of side information constraints? For the purpose of learning dynamical systems, arguably the most relevant notion of "closeness" between two vector fields is one that measures how differently their corresponding trajectories can behave when started from the same initial condition. More formally, we fix a compact set $\Omega \subset \mathbb{R}^n$ and a time horizon T, and we define the following notion of distance between any two vector fields $f, g \in C^\circ_1(\Omega)$:

$$d_{\Omega,T}(f,g) := \sup_{(t,\mathbf{x}_{\text{init}})\in\mathcal{S}} \max\left\{\|\mathbf{x}_f(t,\mathbf{x}_{\text{init}}) - x_g(t,\mathbf{x}_{\text{init}})\|_2, \|\dot{\mathbf{x}}_f(t,\mathbf{x}_{\text{init}}) - \dot{x}_g(t,\mathbf{x}_{\text{init}})\|_2\right\}, \tag{6.27}$$

where $\mathbf{x}_f(t,\mathbf{x}_{\text{init}})$ (resp. $x_g(t,\mathbf{x}_{\text{init}})$) is the trajectory starting from $\mathbf{x}_{\text{init}} \in \Omega$ and following the dynamics of f (resp. g), and

$$\mathcal{S} := \{(t,\mathbf{x}_{\text{init}}) \in [0,T]\times\Omega \mid \mathbf{x}_f(s,\mathbf{x}_{\text{init}}), x_g(s,\mathbf{x}_{\text{init}}) \in \Omega \ \forall s \in [0,t]\}. \tag{6.28}$$

The reason why in the definition of $d_{\Omega,T}$, we take the supremum over $\mathcal{S}$ instead of over $[0,T]\times\Omega$ is to ensure that the trajectories that appear in (6.27) are well defined.

In Section 6.5.1, we show that under some assumptions that are often met in practice, polynomial vector fields can be made arbitrarily close to any continuously-differentiable vector field f (in the sense of (6.27)), even if they are required to satisfy one side information constraint that f is known to satisfy. In Section 6.5.2, we drop our assumptions and generalize this approximation result to any list of side information constraints at the price of allowing an arbitrarily small error in the satisfaction of these constraints. Furthermore, we show that the approximate satisfaction of side information can be certified by a sum of squares proof.

6.5.1 Approximating a vector field while (exactly) satisfying one side information constraint

The following theorem is the main result of this section. We will need the following definition for a subcase of this theorem: Given a collection of sets $A_1, \ldots, A_r$, we define $\mathcal{G}(A_1, \ldots, A_r)$ to be the graph on r vertices labeled by the sets $A_1, \ldots, A_r$, where two vertices A_i and A_j are connected if $A_i \cap A_j \neq \emptyset$.

Theorem 6.5.1. *For any compact set $\Omega \subset \mathbb{R}^n$, time horizon $T > 0$, desired accuracy $\varepsilon > 0$, and vector field $f \in C_1^\circ(\Omega)$ which satisfies one of the following side information constraints (see Section 6.2):*

(i) **Interp**$(\{(\mathbf{x}_i, y_i)\}_{i=1}^m)$*, where $x_1, \ldots, x_m \in \Omega$,*

(ii) **Sym**$(G, \sigma, \rho,)$*,*

(iii) **Pos**$(\{(P_i, N_i)\}_{i=1}^n)$*, where for each $i \in \{1, \ldots, n\}$, $P_i \cap N_i = \emptyset$,*

(iv) **Mon**$(\{(P_{ij}, N_{ij})\}_{i,j=1}^n)$*, where for each $i, j \in \{1, \ldots, n\}$, the sets P_{ij} and N_{ij} belong to different connected components of the graph $\mathcal{G}(P_{i1}, N_{i1}, \ldots, P_{in}, N_{in})$,*

(v) **Inv**$(\{B_i\}_{i=1}^r)$*, where the sets B_i are pairwise nonintersecting, and defined as $B_i := \{x \in \mathbb{R}^n \mid h_{ij}(\mathbf{x}) \geq 0,\ j = 1, \ldots, m_i\}$ for some concave continuously-differentiable functions h_{ij} that satisfy*

$$\forall i \in \{1, \ldots, r\},\ \exists x^i \in B_i \text{ such that } h_{ij}(x^i) > 0 \text{ for } j = 1, \ldots, m_i,$$

(vi) **Grad***,*

(vi') **Ham***,*

there exists a polynomial vector field $p : \mathbb{R}^n \to \mathbb{R}^n$ that satisfies the same side information constraint as f and has $d_{\Omega,T}(f, p) \leq \varepsilon$.

Before we present the proof, we recall the classical Stone-Weierstrass approximation theorem. Note that while the theorem is stated here for scalar-valued functions, it readily extends to vector-valued ones.

Theorem 6.5.2. *(see, e.g., [194]) For any compact set $\Omega \subset \mathbb{R}^n$, scalar $\varepsilon > 0$, and continuous function $f : \Omega \to \mathbb{R}$, there exists a polynomial $p : \mathbb{R}^n \to \mathbb{R}$ such that*

$$\max_{x \in \Omega} |f(x) - p(x)| \leq \varepsilon.$$

For two vector fields $f, g : \mathbb{R}^n \to \mathbb{R}^n$ and a set $\Omega \subseteq \mathbb{R}^n$, let us define

$$\|f - g\|_\Omega := \max_{\mathbf{x} \in \Omega} \|f(\mathbf{x}) - g(\mathbf{x})\|_2.$$

The following proposition relates this quantity to the notion of distance $d_{\Omega,T}(f,g)$ defined in (6.27). We recall that for a scalar $L \geq 0$, a vector field f is said to be L-Lipschitz over Ω if

$$\|f(x) - f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \Omega.$$

Note that any vector field $f \in C_1^\circ(\Omega)$ is L-Lipschitz over a compact set Ω for some nonnegative scalar L.

Proposition 6.5.3. *For any compact set $\Omega \subset \mathbb{R}^n$, time horizon $T > 0$, and two vector fields $f, g \in C_1^\circ(\Omega)$, we have*

$$\|f - g\|_\Omega \leq d_{\Omega,T}(f,g) \leq \max\{Te^{LT}, 1 + LTe^{LT}\}\|f - g\|_\Omega,$$

where $L \geq 0$ is any scalar for which either f or g is L-Lipschitz over Ω.

To present the proof of this proposition, we need to recall the Grönwall-Bellman inequality.

Lemma 6.5.4 (Grönwall-Bellman inequality [35, 112])**.** *Let $I = [a,b]$ denote a nonempty interval on the real line. Let u, α, $\beta : I \to \mathbb{R}$ be continuous functions satisfying*

$$u(t) \leq \alpha(t) + \int_a^t \beta(s)u(s)\,\mathrm{d}s \quad \forall t \in I.$$

If α is nondecreasing and β is nonnegative, then

$$u(t) \leq \alpha(t)e^{\int_a^t \beta(s)\,\mathrm{d}s} \quad \forall t \in I.$$

Proof of proposition 6.5.3. We fix a compact set $\Omega \subset \mathbb{R}^n$, a time horizon $T > 0$, and vector fields $f, g \in C_1^\circ(\Omega)$, with f being L-Lipschitz over Ω for some scalar $L \geq 0$. To see that the first inequality holds, note that for any $x_{\text{init}} \in \Omega$, $\dot{\mathbf{x}}_f(0, \mathbf{x}_{\text{init}}) = f(\mathbf{x}_{\text{init}})$ and $\dot{x}_g(0, \mathbf{x}_{\text{init}}) = g(\mathbf{x}_{\text{init}})$. Therefore, $\|f - g\|_\Omega \leq d_{\Omega,T}(f,g)$.

For the second inequality, fix $(t, x_{\text{init}}) \in \mathcal{S}$, where $\mathcal{S}$ is defined in (6.28). Let us first bound $\|x_f(t, x_{\text{init}}) - x_g(t, x_{\text{init}})\|_2$. By definition of x_f and x_g, we have

$$\begin{aligned} x_f(t, x_{\text{init}}) - x_g(t, x_{\text{init}}) &= \int_0^t f(x_f(s, x_{\text{init}})) - g(x_g(s, x_{\text{init}}))\mathrm{d}s \\ &= \int_0^t f(x_f(s, x_{\text{init}})) - f(x_g(s, x_{\text{init}}))\,\mathrm{d}s \\ &+ \int_0^t f(x_g(s, x_{\text{init}})) - g(x_g(s, x_{\text{init}}))\,\mathrm{d}s. \end{aligned}$$

Using the triangular inequality, we get

$$\begin{aligned}\|x_f(t, x_{\text{init}}) - x_g(t, x_{\text{init}})\|_2 \leq &\int_0^t \|f(x_f(s, x_{\text{init}})) - f(x_g(s, x_{\text{init}}))\|_2 \, \mathrm{d}s \\ &+ \int_0^t \|f(x_g(s, x_{\text{init}})) - g(x_g(s, x_{\text{init}}))\|_2 \, \mathrm{d}s.\end{aligned} \tag{6.29}$$

Because the function f is L-Lipschitz over Ω, we have

$$\|f(x_f(s, x_{\text{init}})) - f(x_g(s, x_{\text{init}}))\|_2 \leq L\|x_f(s, x_{\text{init}}) - x_g(s, x_{\text{init}})\|_2 \quad \forall s \in [0, t].$$

Furthermore, we know that for all $s \in [0, t]$, $x_g(s, x_{\text{init}}) \in \Omega$, and therefore

$$\|f(x_g(s, x_{\text{init}})) - g(x_g(s, x_{\text{init}}))\|_2 \leq \|f - g\|_\Omega.$$

We can hence further bound the left hand side of (6.29) as

$$\|x_f(t, x_{\text{init}}) - x_g(t, x_{\text{init}})\|_2 \leq L \int_0^t \|x_f(s, x_{\text{init}}) - x_g(s, x_{\text{init}})\|_2 \, \mathrm{d}s + t\|f - g\|_\Omega.$$

By Lemma 6.5.4, we get

$$\|x_f(t, x_{\text{init}}) - x_g(t, x_{\text{init}})\|_2 \leq te^{Lt}\|f - g\|_\Omega. \tag{6.30}$$

Next, we bound the quantity$\|\dot{x}_f(t, x_{\text{init}}) - \dot{x}_g(t, x_{\text{init}})\|_2$, which can be expressed in terms of the vector fields f and g as $\|f(x_f(t, x_{\text{init}})) - g(x_g(t, x_{\text{init}}))\|_2$. We have

$$\begin{aligned}\|f(x_f(t, x_{\text{init}})) - g(x_g(t, x_{\text{init}}))\|_2 &\leq \|f(x_f(t, x_{\text{init}})) - f(x_g(t, x_{\text{init}}))\|_2 \\ &\quad + \|f(x_g(t, x_{\text{init}})) - g(x_g(t, x_{\text{init}}))\|_2 \\ &\leq L\|x_f(t, x_{\text{init}}) - x_g(t, x_{\text{init}})\|_2 + \|f - g\|_\Omega \\ &\leq (1 + Lte^{Lt})\|f - g\|_\Omega,\end{aligned} \tag{6.31}$$

where the first inequality follows from the triangular inequality, the second from the definition of $\|\cdot\|_\Omega$ and the fact that f is L-Lipschitz over Ω, and the third one from (6.30).

Putting (6.30) and (6.31) together, and using the fact that $t \leq T$, we have

$$\begin{aligned}&\max\left\{\|x_f(t, x_{\text{init}}) - x_g(t, x_{\text{init}})\|_2, \|f(x_f(t, x_{\text{init}})) - g(x_g(t, x_{\text{init}}))\|_2\right\} \\ &\qquad \leq \max\{te^{Lt}, 1 + Lte^{Lt}\}\|f - g\|_\Omega \\ &\qquad \leq \max\{Te^{LT}, 1 + LTe^{LT}\}\|f - g\|_\Omega.\end{aligned}$$

Taking the supremum over $(t, x_{\text{init}}) \in \mathcal{S}$, we get

$$d_{\Omega,T}(f, g) \leq \max\{Te^{LT}, 1 + LTe^{LT}\}\|f - g\|_\Omega.$$

□

Proof of Theorem 6.5.1. Let us fix a compact set $\Omega \subset \mathbb{R}^n$, a time horizon $T > 0$, and a desired accuracy $\varepsilon > 0$. Let $f \in C_1^\circ(\Omega)$ be a vector field that satisfies any one of the side information constraints stated in the theorem. Note that f is L-Lipschitz over Ω for some $L \geq 0$. We claim that for any $\delta > 0$, there exists a polynomial vector field $p : \mathbb{R}^n \to \mathbb{R}^n$ that satisfies the same side information constraint as f and the inequality

$$\|f - p\|_\Omega \leq \delta.$$

By Proposition 6.5.3, if we take

$$\delta = \varepsilon / \max\{Te^{LT}, 1 + LTe^{LT}\}, \tag{6.32}$$

this shows that there exists a polynomial vector field $p : \mathbb{R}^n \to \mathbb{R}^n$ that satisfies the same side information as f and the inequality

$$d_{\Omega,T}(f, p) \leq \varepsilon.$$

We now give a case-by-case proof of our claim above depending on which side information f satisfies. Throughout the rest of the proof, the constant δ is fixed as in (6.32).

- **Case (i):** Suppose $f \in \textbf{Interp}(\{(\mathbf{x}_i, y_i)\}_{i=1}^m)$, where $x_1, \ldots, x_m \in \Omega$. Without loss of generality, we assume that the points x_i are all different, or else we can discard the redundant constraints. Let δ' be a positive constant that will be fixed later. By theorem 6.5.2, there exists a polynomial vector field q that satisfies $\|f - q\|_\Omega \leq \delta'$. We claim that there exists a polynomial $\tilde{q}$ of degree $m - 1$ such that

$$(q + \tilde{q})(x_i) = y_i \quad i = 1, \ldots, m, \tag{6.33}$$

and $\|\tilde{q}\|_\Omega \leq C\delta'$, where C is a constant depending only on the points x_i and the set Ω. Indeed, (6.33) can be viewed as a linear system of equations where the unknowns are the coefficients of $\tilde{q}$ in some basis. For example, if we let $N = \binom{n+m-1}{n}$ and $\tilde{q}_{\text{coeff}} \in \mathbb{R}^{N \times n}$ be the matrix whose j-th column is the vector of coefficients of $\tilde{q}_j$ in the standard monomial basis, then (6.33) can be written as

$$A\, \tilde{q}_{\text{coeff}} = \Delta, \tag{6.34}$$

where $\Delta \in \mathbb{R}^{m \times n}$ is the matrix whose i-th row is given by $y_i^T - q(x_i)^T$, and $A \in \mathbb{R}^{m \times N}$ is the matrix whose i-th row is the vector of all standard monomials in n variables and of degree up to $m - 1$ evaluated at the point x_i. One can verify that the rows of the matrix A are linearly independent (see, e.g., [61, Corollary 4.4]), and so the matrix AA^T is invertible. If we let $A^+ = A^T(AA^T)^{-1}$, then $\tilde{q}_{\text{coeff}} = A^+\Delta$ is a solution to (6.34). Since the matrix A^+ only depends on the points x_i, and since all the entries of the matrix Δ are bounded in absolute value by δ', there exists a constant c such that all the entries of the matrix $\tilde{q}_{\text{coeff}}$ are bounded in absolute value by $c\delta'$. Since the set Ω is compact, there exists a constant C depending only on the points x_i and the set Ω such that $\|\tilde{q}\|_\Omega \leq C\delta'$.

Finally, by taking $p := q + \tilde{q}$, we get $p \in \mathbf{Interp}(\{(\mathbf{x}_i, y_i)\}_{i=1}^m)$, and

$$\|f - p\|_\Omega \leq \delta'(1 + C).$$

We take $\delta' = \frac{\delta}{1+C}$ to conclude the proof for this case.

· **Case (ii):** Suppose $f \in \mathbf{Sym}(G, \sigma, \rho,)$, where G is a finite group. Let δ' be a positive constant that will be fixed later. By theorem 6.5.2, there exists a polynomial vector field p that satisfies $\|f - p\|_\Omega \leq \delta'$. Let $p^G : \mathbb{R}^n \to \mathbb{R}^n$ be the polynomial defined as

$$p^G(x) := \frac{1}{|G|} \sum_{g \in G} \rho(g^{-1}) p(\sigma(g)x) \quad \forall x \in \mathbb{R}^n,$$

where $|G|$ is the size of the group G. We claim that $p^G \in \mathbf{Sym}(G, \sigma, \rho,)$. Indeed, for any $h \in G$ and $x \in \Omega$, using the fact that σ is a group homomorphism, we get

$$p^G(\sigma(h)x) = \frac{1}{|G|} \sum_{g \in G} \rho(g^{-1}) p(\sigma(gh)x).$$

By doing the change of variables $g' = gh$ in the sum above, and using the fact that ρ is a group homomorphism, we get

$$\begin{aligned} p^G(\sigma(h)x) &= \frac{1}{|G|} \sum_{g' \in G} \rho(hg'^{-1}) p(\sigma(g')x) \\ &= \frac{1}{|G|} \sum_{g' \in G} \rho(h)\rho(g'^{-1}) p(\sigma(g')x). \\ &= \rho(h) p^G(x). \end{aligned}$$

We now claim that by by taking $\delta' = \delta \left(\frac{1}{|G|} \sum_{g \in G} \|\rho(g^{-1})\| \right)^{-1}$, where $\|\cdot\|$ denotes the operator norm of its matrix argument, we get $\|f - p^G\|_\Omega \leq \delta$. Indeed,

$$\begin{aligned} f(x) - p^G(x) &= \frac{1}{|G|} \sum_{g \in G} \left(f(x) - \rho(g^{-1}) p(\sigma(g)x) \right) \\ &= \frac{1}{|G|} \sum_{g \in G} \rho(g^{-1}) \left(\rho(g) f(x) - p(\sigma(g)x) \right) \\ &= \frac{1}{|G|} \sum_{g \in G} \rho(g^{-1}) \left(f(\sigma(g)x) - p(\sigma(g)x) \right), \end{aligned}$$

where in the last equation, we used the fact that $f \in \mathbf{Sym}(G, \sigma, \rho,)$. Therefore,

$$\begin{aligned}\|f(x) - p^G(x)\| &\leq \frac{1}{|G|} \sum_{g \in G} \|\rho(g^{-1})\| \|f(\sigma(g)x) - p(\sigma(g)x)\|_2 \\ &\leq \left(\frac{1}{|G|} \sum_{g \in G} \|\rho(g^{-1})\| \right) \delta' = \delta.\end{aligned}$$

· **Case (iii):** If $f \in \mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$, where for each $i \in \{1, \ldots, n\}$, the sets P_i and N_i are subsets of Ω and satisfy $P_i \cap N_i = \emptyset$.

For $i = 1, \ldots, n$, let d_i denote the distance between the sets P_i and N_i:

$$d_i := \min_{x \in P_i, x' \in N_i} \|x - x'\|_2.$$

Since P_i and N_i are compact sets with empty intersection, the scalar d_i is positive. Fix γ to be any positive scalar smaller than $\min_{i=1,\ldots,n} d_i$. For $i = 1, \ldots, n$, let

$$P_i^\gamma := \{x + \frac{\gamma}{2} z \mid x \in P_i, z \in \mathbb{R}^n, \text{ and } \|z\|_2 \leq 1\},$$

$$N_i^\gamma := \{x + \frac{\gamma}{2} z \mid x \in N_i, z \in \mathbb{R}^n, \text{ and } \|z\|_2 \leq 1\}.$$

With our choice of γ, $P_i^\gamma \cap N_i^\gamma = \emptyset$ for $i = 1, \ldots, n$. Let $\psi : \mathbb{R}^n \to \mathbb{R}^n$ be the piecewise-constant function defined as

$$\psi_i(x) = \begin{cases} 1 & \text{if } x \in P_i^\gamma \\ -1 & \text{if } x \in N_i^\gamma \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, \ldots, n,$$

and $\phi^\gamma : \mathbb{R}^n \to \mathbb{R}$ be the "bump-like" function that is equal to $e^{-\frac{1}{1-\|z\|^2}}$ when $\|z\|_2 \leq \frac{\gamma}{2}$ and 0 elsewhere. Let $\psi^{\text{conv}} : \mathbb{R}^n \to \mathbb{R}^n$ be the normalized convolution of ψ with ϕ^γ , i.e.,

$$\psi^{\text{conv}}(x) := \frac{1}{\int_{z \in \mathbb{R}^n} \phi^\gamma(z) \mathrm{d}z} \int_{z \in \mathbb{R}^n} \psi(x + z) \phi^\gamma(z) \mathrm{d}z.$$

Note that ψ^{conv} is a continuous function as it is the convolution of a piecewise-constant function ψ with a continuous function ϕ^γ. Moreover, for each $i \in \{1, \ldots, n\}$, ψ_i^{conv} satisfies

$$\psi_i^{\text{conv}}(x) = 1 \; \forall x \in P_i, \; \psi_i^{\text{conv}}(x) = -1 \; \forall x \in N_i, \text{ and } |\psi_i^{\text{conv}}(x)| \leq 1 \; \forall x \in \Omega.$$

Now let $f^\delta \in C_1^\circ(\Omega)$ be the vector field defined component-wise by

$$f_i^\delta(x) = f_i(x) + \frac{\delta}{2\sqrt{n}} \psi_i^{\text{conv}}(x) \quad i = 1, \ldots, n.$$

Note that for $i = 1, \ldots, n$, the function f_i^δ is continuous, bounded below by $\frac{\delta}{2\sqrt{n}}$ on P_i, and bounded above by $\frac{-\delta}{2\sqrt{n}}$ on N_i. Moreover $\|f - f^\delta\|_\Omega \leq \frac{\delta}{2}$. Theorem 6.5.2 guarantees the existence of a polynomial vector field p such that $\|f^\delta - p\|_\Omega \leq \frac{\delta}{2\sqrt{n}}$. In particular, $p \in \mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$ and satisfies $\|f - p\|_\Omega \leq \delta$.

- **Case (iv):** If $f \in \mathbf{Mon}(\{(P_{ij}, N_{ij})\}_{i,j=1}^n)$, where for each $i, j \in \{1, \ldots, n\}$, the sets P_{ij} and N_{ij} are subsets of Ω and belong to different connected components of the graph $\mathcal{G}(P_{i1}, N_{i1}, \ldots, P_{in}, N_{in})$. (Recall the definition of this graph from the first paragraph of Section 6.5.1.)

Consider an index $i \in \{1, \ldots, n\}$, and let $C_{i1}, \ldots, C_{ir_i}$ be the connected components of the graph $\mathcal{G}(P_{i1}, N_{i1}, \ldots, P_{in}, N_{in})$. Let U_{il} be the union of the sets in component C_{il}. Since the sets $U_{i1}, \ldots, U_{ir_i}$ are compact and pairwise non-intersecting, the minimum distance $d_i := \min_{x \in U_{il}, x' \in U_{il'}, l \neq l'} \|x - x'\|_2$ between any two of them is positive. Fix γ_i to be a positive scalar smaller than d_i, and for each $l \in \{1, \ldots, r_i\}$, let

$$U_{il}^{\gamma_i} := \{x + \frac{\gamma_i}{2} z \mid x \in U_{il}, z \in \mathbb{R}^n, \text{ and } \|z\|_2 \leq 1\}.$$

Define $\psi_i : \mathbb{R}^n \to \mathbb{R}$ to be the piecewise-linear function defined as

$$\psi_i(x) = \begin{cases} \sum_{j: P_{ij} \in C_l} x_j - \sum_{j: N_{ij} \in C_l} x_j & \text{if } x \in U_{il}^{\gamma_i} \text{ for some } l \in \{1, \ldots, r_i\} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\psi_i^{\text{conv}} : \mathbb{R}^n \to \mathbb{R}$ be the normalized convolution of ψ_i with the "bump-like" function $\phi^{\gamma_i} : \mathbb{R}^n \to \mathbb{R}$ that is equal to $e^{-\frac{1}{1-\|z\|^2}}$ when $\|z\|_2 \leq \frac{\gamma_i}{2}$ and 0 elsewhere; that is

$$\psi_i^{\text{conv}}(x) := \frac{1}{\int_{z \in \mathbb{R}^n} \phi^{\gamma_i}(z) \mathrm{d}z} \int_{z \in \mathbb{R}^n} \psi_i(x + z) \phi^{\gamma_i}(z) \mathrm{d}z.$$

The function ψ_i^{conv} is continuously differentiable (because ϕ^{γ_i} is continuously differentiable) and satisfies

$$\frac{\partial \psi_i^{\text{conv}}}{\partial x_j}(x) = 1 \; \forall x \in P_{ij}, \; \frac{\partial \psi_i^{\text{conv}}}{\partial x_j}(x) = -1 \; \forall x \in N_{ij},$$

$$|\psi_i^{\text{conv}}(x)| \leq \sup_{x \in \Omega} |\psi_i(x)| \quad \forall x \in \Omega.$$

Now, let $\psi^{\text{conv}} := (\psi_1^{\text{conv}}, \ldots, \psi_n^{\text{conv}})^T$ and $f^{\delta'}(x) := f(x) + \delta' \psi^{\text{conv}}(x)$ for a constant $\delta' > 0$ that will be fixed later. Note that $\|f^{\delta'} - f\|_\Omega \leq \delta' \|\psi^{\text{conv}}\|_\Omega$, and for each pair of indices $i, j \in \{1, \ldots, n\}$,

$$\frac{\partial f_i^{\delta'}}{\partial x_j}(x) \geq \delta' \; \forall x \in P_{ij}, \; \frac{\partial f_i^{\delta'}}{\partial x_j}(x) \leq -\delta' \; \forall x \in N_{ij}.$$

A generalization of the Stone-Weierstrass approximation result stated in Theorem 6.5.2 to continuously-differentiable functions (see, e.g., [167]) guarantees the

existence of a polynomial $p : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\|f^{\delta'} - p\|_\Omega \leq \delta', \quad \sup_{x\in\Omega} \left| \frac{\partial f_i^{\delta'}}{\partial x_j}(x) - \frac{\partial p_i}{\partial x_j}(x) \right| \leq \delta'/2 \quad \forall i, j \in \{1, \ldots, n\}.$$

In particular, $p \in \mathbf{Mon}(\{P_{ij}, N_{ij}\}_{i,j=1}^n)$ and satisfies $\|f - p\|_\Omega \leq \delta'(1 + \|\psi^{\text{conv}}\|_\Omega)$. We conclude the proof by taking $\delta' = \frac{\delta}{1+\|\psi^{\text{conv}}\|_\Omega}$.

• **Case (v):** If $f \in \mathbf{Inv}(\{B_i\}_{i=1}^r)$, where the sets B_i are subsets of Ω, pairwise nonintersecting, and defined as $B_i := \{x \in \mathbb{R}^n \mid h_{ij}(\mathbf{x}) \geq 0,\ j = 1, \ldots, m_i\}$ for some continuously-differentiable concave functions $h_{ij} : \mathbb{R}^n \to \mathbb{R}$ that satisfy

$$\forall i \in \{1, \ldots, r\},\ \exists x^i \in B_i \text{ such that } h_{ij}(x^i) > 0 \text{ for } j = 1, \ldots, m_i. \tag{6.35}$$

By the same argument as that for **Case (iii)**, for each $i \in \{1, \ldots, r\}$, there exists a continuous function $\psi_i^{\text{conv}} : \mathbb{R}^n \to \mathbb{R}$ that satisfies

$$\psi_i^{\text{conv}}(x) = 1\ \forall x \in B_i, \psi_i^{\text{conv}}(x) = 0\ \forall x \in \cup_{i'\neq i} B_{i'},\ |\psi_i^{\text{conv}}(x)| \leq 1\ \forall x \in \Omega.$$

Let $\delta' := \frac{\delta}{2r(1+\max_{x,x'\in\Omega}\|x-x'\|_2)}$, and for $i = 1 \ldots, r$, let $x^i \in B_i$ be any point satisfying $h_{ij}(x^i) > 0$ for $j = 1, \ldots, m_i$. Consider the continuous vector field

$$f^{\delta'}(x) := f(x) - \delta' \sum_{i=1}^r \psi_i^{\text{conv}}(x)(x - x^i).$$

For every $x \in \Omega$, the triangular inequality gives

$$\begin{aligned} \|f(x) - f^{\delta'}(x)\|_2 &\leq \delta' \sum_{i=1}^r \|x - x^i\|_2 \\ &\leq r\delta' \max_{x'\in\Omega} \|x - x'\|_2 = \frac{\delta}{2}, \end{aligned}$$

and so $\|f - f^{\delta'}\|_\Omega \leq \delta/2$. Furthermore, for each $i \in \{1, \ldots, r\}$, for each $j \in \{1, \ldots, m_i\}$, and for each $x \in B_i$ satisfying $h_{ij}(x) = 0$,

$$\begin{aligned} \langle f^{\delta'}(x), \nabla h_{ij}(x)\rangle &= \langle f(x), \nabla h_{ij}(x)\rangle - \delta' \sum_{k=1}^r \phi_k^{\text{conv}}(x)\langle x - x^k, \nabla h_{ij}(x)\rangle \\ &= \langle f(x), \nabla h_{ij}(x)\rangle - \delta'\langle x - x^i, \nabla h_{ij}(x)\rangle \\ &\geq -\delta'\langle x - x^i, \nabla h_{ij}(x)\rangle \\ &\geq \delta'(h_{ij}(x^i) - h_{ij}(x)) \\ &= \delta' h_{ij}(x^i), \end{aligned} \tag{6.36}$$

where the second equality follows from the definition of ψ_i^{conv} and the fact that $x \in B_i$, the first inequality from the fact that$f \in \mathbf{Inv}(\{B_i\}_{i=1}^r)$, the second inequality from concavity of the function h_{ij}, and the last equality from the fact that $h_{ij}(x) = 0$.

For a constant $\delta'' > 0$ that will be fixed later, Theorem 6.5.2 guarantees the existence of a polynomial vector field p such that $\|f^{\delta'} - p\|_\Omega \leq \delta''$. By triangular inequality we have $\|f - p\|_\Omega \leq \delta/2 + \delta''$. Furthermore, for each $i \in \{1, \ldots, r\}$, for each $j \in \{1, \ldots, m_i\}$, and for each $x \in B_i$ satisfying $h_{ij}(x) = 0$, we have

$$\begin{aligned}\langle p(x), \nabla h_{ij}(x)\rangle &= \langle p(x) - f^{\delta'}(x), \nabla h_{ij}(x)\rangle + \langle f^{\delta'}(x), \nabla h_{ij}(x)\rangle \\ &\geq -\delta''\|\nabla h_{ij}(x)\|_2 + \delta' h_{ij}(x^i)\end{aligned}$$

due to (6.36) and the Cauchy-Schwarz inequality. Let

$$\delta'' := \min\left\{\frac{\delta}{2}, \min_{i\in\{1,\ldots,r\}} \min_{j\in\{1,\ldots,m_i\}, x\in B_i} \delta' \frac{h_{ij}(x^i)}{\|\nabla h_{ij}(x)\|_2}\right\},$$

and note that $\delta'' > 0$ as we needed before because $h_{ij}(x^i) > 0$ for each $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, m_i\}$. With this choice of δ'', we get that $p \in \mathbf{Inv}(\{B_i\}_{i=1}^r)$ and $\|f - p\|_\Omega \leq \delta$.

- **Case (vi):** If $f \in \mathbf{Grad}$. In this case, there exists a continuously-differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ such that $f(\mathbf{x}) = -\nabla V(\mathbf{x})$. A generalization of the Stone-Weierstrass theorem to continuously-differentiable functions (see, e.g., [167]) guarantees the existence of a polynomial $W : \mathbb{R}^n \to \mathbb{R}$ such that

$$\max_{x\in\Omega} \|\nabla V(x) - \nabla W(x)\|_2 \leq \delta.$$

Letting $p(x) = -\nabla W(x)$, we get that $p \in \mathbf{Grad}$ and $\|f - p\|_\Omega \leq \delta$.

- **Case (vi'):** If $f \in \mathbf{Ham}$. The proof for this case is analogous to **Case (vi)**. □

6.5.2 Approximating a vector field while approximately satisfying multiple side information constraints

It is natural to ask whether theorem 6.5.1 can be generalized to allow for polynomial approximation of vector fields satisfying *multiple* side information constraints. It turns out that our proof idea of "smoothing by convolution" can be used to show that the answer is positive if the following three conditions hold: (i) the side information constraints are of type $\mathbf{Interp}, \mathbf{Pos}, \mathbf{Mon}, \mathbf{Inv}$, or $\mathbf{Sym}$, (ii) each side information constraint satisfies the assumptions of Theorem 6.5.1, and (iii) the regions of the space where the first four types of side information constraints are imposed are pairwise nonintersecting. In absence of condition (iii), the answer is no longer positive as the next example shows.

Example 4. *Consider the univariate vector field $f : \mathbb{R} \to \mathbb{R}$ given by*

$$f(x) := \begin{cases} 0 & x \geq 0 \\ -e^{-\frac{1}{x^2}} & x < 0. \end{cases}$$

Side information S	**Functional** $L_{S,\Omega}(f)$
$\mathbf{Interp}(\{(\mathbf{x}_i, y_i)\})_{i=1}^m)$ with $x_i \in \Omega$ for $i = 1 \ldots, m$	$\max\limits_{i=1,\ldots,m} \|f(x_i) - y_i\|_2$
$\mathbf{Sym}(G, \sigma, \rho,)$	$\max\limits_{g\in G} \max\limits_{\substack{i=1,\ldots,n \\ x\in\Omega}} \|f_i(\sigma(g)x) - (\rho(g)f(x))_i\|$
$\mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$ with $P_i, N_i \subseteq \Omega$ for $i = 1, \ldots, n$	$\max\limits_{i=1,\ldots,n} \max\left\{0, \max\limits_{x\in P_i} -f_i(x), \max\limits_{x\in N_i} f_i(x)\right\}$
$\mathbf{Mon}(\{(P_{ij}, N_{ij})\}_{i,j=1}^n)$ with $P_{ij}, N_{ij} \subseteq \Omega$ for $i, j = 1, \ldots, n$	$\max\limits_{i,j=1,\ldots,n} \max\left\{0, \max\limits_{x\in P_{ij}} -\dfrac{\partial f_i}{\partial x_j}(x), \max\limits_{x\in N_{ij}} \dfrac{\partial f_i}{\partial x_j}(x)\right\}$
$\mathbf{Inv}(\{B_i\}_{i=1}^r)$ where $B_i := \{x \mid h_{ij}(x) \geq 0$ $\forall j \in \{1, \ldots, m_i\}\} \subseteq \Omega$ for $i = 1, \ldots, r$	$\max\limits_{i=1\ldots,r} \max\limits_{\substack{\mathbf{x}\in B_i \\ j\in\{1,\ldots,m_i\} \\ h_{ij}(\mathbf{x})=0}} \max\{0, -\langle f(\mathbf{x}), \nabla h_{ij}(\mathbf{x})\rangle\}$
Grad	$\inf\limits_{V:\mathbb{R}^n\to\mathbb{R}} \max\limits_{\substack{i=1,\ldots,n \\ x\in\Omega}} \left\lvert f_i(x) + \dfrac{\partial V}{\partial x_i}(x)\right\rvert$
Ham	$\inf\limits_{H:\mathbb{R}^n\to\mathbb{R}} \max\limits_{\substack{(p,q)\in\Omega, \\ i=1\ldots,n/2}} \max\left\{\left\lvert f_i(p,q) + \dfrac{\partial H}{\partial q_i}(p,q)\right\rvert, \left\lvert f_{i+n/2}(p,q) - \frac{\partial H}{\partial p_i}(p,q)\right\rvert\right\}$

Table 6.2: For each side information S, the functional $L_{S,\Omega} : C_1^\circ(\Omega) \to \mathbb{R}$ quantifies how close a vector field $f \in C_1^\circ(\Omega)$ is to satisfying S.

This vector field is continuously differentiable over $\mathbb{R}$ *and satisfies the following combination of side information constraints:*

$$\mathbf{Interp}(\{(0,0),(1,0)\}) \text{ and } \mathbf{Mon}(\{([-1,1],\emptyset)\}). \tag{6.37}$$

In other words, f *is nondecreasing on the interval* $[-1,1]$ *and satisfies* $f(0) = f(1) = 0$*. Yet, the only polynomial vector field that satisfies the constraints in eq. (6.37) is the identically zero polynomial. As a result, the vector field* f *cannot be approximated arbitrarily well over* $[-1,1]$ *by polynomial vector fields that satisfy the side information constraints in eq. (6.37).*

To overcome difficulties associated with such examples, we introduce the notion of approximate satisfiability of side information over a compact set $\Omega \subset \mathbb{R}^n$. Before we give a formal definition of this notion, for each side information constraint S, we present in Table 6.2 a functional $L_{S,\Omega} : C_1^\circ(\Omega) \to \mathbb{R}$ that measures how close a vector field in $C_1^\circ(\Omega)$ is to satisfying the side information S. One can verify that the functional $L_{S,\Omega}$ has the following two properties: (i) for any vector field $f \in C_1^\circ(\Omega)$,

$$L_{S,\Omega}(f) = 0 \text{ if and only if } f \text{ satisfies } S, \tag{6.38}$$

and (ii) for any $\delta > 0$, there exists $\gamma > 0$, such that for any two vector fields $f, \hat{f} \in C_1^\circ(\Omega)$,

$$\|f - \hat{f}\|_\Omega \leq \gamma \text{ and } \max_{\substack{x\in\Omega,\\ i,j=1,\dots,n}} \left| \frac{\partial f_i}{\partial x_j}(x) - \frac{\partial \hat{f}_i}{\partial x_j}(x) \right| \leq \gamma \implies |L_{S,\Omega}(f) - L_{S,\Omega}(\hat{f})| \leq \delta. \tag{6.39}$$

Indeed, take e.g. $S = \mathbf{Inv}(\{B_i\}_{i=1}^r)$, where $B_i := \{x \in \mathbb{R}^n \mid h_{ij}(x) \geq 0,\ j = 1, \dots, m_i\}$. It is clear from condition eq. (6.7) that $L_{S,\Omega}(f) = 0$ if and only if $f \in \mathbf{Inv}(\{B_i\}_{i=1}^r)$. To verify the second property, let $\delta > 0$ be given. If we take

$$\gamma = \delta \left\{ \max_{\substack{x\in\Omega,\\ i=1\dots,r\\ j=1\dots,m_i}} \|\nabla h_{ij}(x)\| \right\}^{-1},$$

it is easy to see that for any two vector fields $f, \hat{f} \in C_1^\circ(\Omega)$ satisfying $\|f - \hat{f}\|_\Omega \leq \gamma$, we must have $|L_S(f) - L_S(\hat{f})| \leq \delta$. Indeed, let $i \in \{1, \dots, r\}$ and $x \in B_i$ be such that $h_{ij}(x) = 0$ for some $j \in \{1, \dots, m_i\}$. Then, the Cauchy-Schwarz inequality and our choice of γ give

$$|\langle f(x), \nabla h_{ij}(x)\rangle - \langle \hat{f}(x), \nabla h_{ij}(x)\rangle| \leq \|f - \hat{f}\|_\Omega \|\nabla h_{ij}(x)\| \leq \delta.$$

The desired result follows by taking the maximum over i, j, and x.

Definition 6.5.5 (δ-satisfiability)**.** *Let $\Omega \subset \mathbb{R}^n$ be a compact set and consider any side information S presented in Table 6.2 together with its corresponding functional $L_{S,\Omega}$. For a scalar $\delta > 0$, we say that a vector field $f \in C_1^\circ(\Omega)$ δ-satisfies S if $L_{S,\Omega}(f) \leq \delta$.*

From a practical standpoint, for small values of δ, it is reasonable to substitute the requirement of exact satisfiability of side information for δ-satisfiability. This is especially true since most optimization solvers return an approximate numerical solution anyway. The following theorem shows that polynomial vector fields can approximate any continuously-differentiable vector field f and satisfy the same side information as f up to an arbitrarily small error tolerance δ. It also shows that in the context of learning a vector field from trajectory data, one can always impose δ-satisfiability on a candidate polynomial vector field via semidefinite programming.

Theorem 6.5.6. *For any compact set $\Omega \subset \mathbb{R}^n$, time horizon $T > 0$, desired approximation accuracy $\varepsilon > 0$, desired side information satisfiability accuracy $\delta > 0$, and for any vector field $f \in C_1^\circ(\Omega)$ that satisfies any combination of the side information constraints from the first column of Table 6.2, there exists a polynomial vector field $p : \mathbb{R}^n \to \mathbb{R}^n$ that δ-satisfies the same combination of side information as f and has $d_{\Omega,T}(f, p) \leq \varepsilon$.*

Moreover, if the set Ω, the sets P_i, N_i in the definition of $\mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$, the sets P_{ij}, N_{ij} in the definition of $\mathbf{Mon}(\{P_{ij}, N_{ij}\}_{i,j=1}^n)$, and the sets B_i in the definition

of $\mathbf{Inv}(\{B_i\}_{i=1}^r)$*) are all closed basic semialgebraic and their defining polynomials satisfy the Archimedian property, then* δ*-satisfiability of all side information constraints by the polynomial vector field* p *has a sum of squares certificate.*

Proof. Let $f \in C_1^\circ(\Omega)$ satisfy a list $S_1, \ldots, S_k$ of side information constraints from the first column of Table 6.2, and let the scalars $T, \varepsilon, \delta > 0$ be fixed. A generalization of the Stone-Weierstrass approximation theorem to continuously-differentiable functions (see, e.g., [167]) guarantees that for any $\gamma > 0$, there exists a polynomial $p^\gamma : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\|f - p^\gamma\|_\Omega \leq \gamma \text{ and } \max_{\substack{x\in\Omega \\ i,j=1,\ldots,n}} \left| \frac{\partial f_i}{\partial x_j}(x) - \frac{\partial p_i^\gamma}{\partial x_j}(x) \right| \leq \gamma. \tag{6.40}$$

For the rest of this paragraph, for any $\gamma > 0$, we fix an (arbitrary) choice for the polynomial p^γ. Since for each $i \in \{1, \ldots, k\}$, the functional $L_{S_i,\Omega}$ satisfies (6.39), there exists a scalar $\gamma_i > 0$ for which $L_{S_i,\Omega}(p^\gamma) \leq \delta/2$ for any $\gamma \in (0, \gamma_i]$. If we let

$$\bar{\gamma} := \min\{\varepsilon / \max\{Te^{LT}, 1 + LTe^{LT}\}, \gamma_1, \ldots, \gamma_k\},$$

where $L > 0$ is any scalar for which f is L-Lipschitz over Ω, then the polynomial $p := p^{\bar{\gamma}}$ $\delta/2$-satisfies $S_1, \ldots, S_k$ (and hence δ-satisfies $S_1, \ldots, S_k$), and because of Proposition 6.5.3, also satisfies $d_{\Omega,T}(f, p) \leq \varepsilon$.

To prove the second claim of the theorem, observe that for each $\ell \in \{1, \ldots, k\}$, the fact that p $\delta/2$-satisfies S_ℓ implies the following inequalities:[8]

- If $S_\ell = \mathbf{Sym}(G, \sigma, \rho,)$,

  $$p_i(\sigma(g)x) - (\rho(g)p(x))_i + \delta > 0 \text{ and } (\rho(g)p(x))_i - p_i(\sigma(g)x) + \delta > 0 \; \forall x \in \Omega,$$

  for $g \in G$ and $i = 1, \ldots, n$;

- If $S_\ell = \mathbf{Pos}(\{(P_i, N_i)\}_{i=1}^n)$,

  $$p_i(x) + \delta > 0 \; \forall x \in P_i \text{ and } -p_i(x) + \delta > 0 \; \forall x \in N_i,$$

  for $i = 1, \ldots, n$;

- If $S_\ell = \mathbf{Mon}(\{(P_{ij}, N_{ij})\}_{i,j=1}^n)$,

  $$\frac{\partial p_i}{\partial x_j}(x) + \delta > 0 \; \forall x \in P_{ij}, \text{ and } -\frac{\partial p_i}{\partial x_j}(x) + \delta > 0 \; \forall x \in N_{ij},$$

  for $i, j = 1, \ldots, n$;

- If $S_\ell = \mathbf{Inv}(\{B_i\}_{i=1}^r)$,

  $$\langle p(\mathbf{x}), \nabla h_{ij}(\mathbf{x}) \rangle + \delta > 0 \; \forall x \in B_i \cap \{x \in \mathbb{R}^n \mid h_{ij}(\mathbf{x}) = 0\},$$

[8] We exclude the case $S_\ell = \mathbf{Interp}(\{(\mathbf{x}_i, y_i)\}_{i=1}^m)$ because verifying δ-satisfiability is trivial there, and the case $S_\ell = \mathbf{Ham}$ because the argument for it is similar to that of $S_\ell = \mathbf{Grad}$.

for $i = 1 \ldots, r$, $j = 1, \ldots, m_i$;

- If $S_\ell = \textbf{Grad}$,

$$p_i(x) + \frac{\partial V}{\partial x_i}(x) + \delta > 0 \text{ and } -p_i(x) - \frac{\partial V}{\partial x_i}(x) + \delta > 0 \quad \forall x \in \Omega,$$

for $i = 1 \ldots, n$, where $V : \mathbb{R}^n \to \mathbb{R}$ is a polynomial function. (The fact that V can be taken to be a polynomial function follows from an other application of the generalization of Stone-Weierstrass approximation theorem that was used at the beginning of this proof.)

Observe that each of the above inequalities states that a certain polynomial is positive over a certain closed basic semialgebraic set whose defining polynomials satisfy the Archimedian property by assumption. Therefore, by Putinar's Positivstellesatz (Theorem 6.3.1), there exists a nonnegative integer d such that each one of these inequalities has a degree-d sos-certificate (see eq. (6.11)). Therefore, δ-satisfiability of each side information $S_1, \ldots, S_k$ by the vector field p can be proven by a sum of squares certificate. □

6.6 Discussion and future research directions

From a computational perspective, our approach to learning dynamical systems from trajectory data while leveraging side information relies on convex optimization. If the side information of interest is **Interp**, **Sym**, **Grad**, or **Ham**, then our approach leads to a least-squares problem, and thus can be implemented at large scale. For side information constraints of **Pos**, **Mon**, or **Inv**, our approach requires solutions to semidefinite programs. Classical interior-point methods for SDP come with polynomial-time solvability guarantees (see e.g. [202]), and in practice scale to problems of moderate sizes. In the field of dynamical systems, many applications of interest involve a limited number of state variables, and therefore our approach to learning such systems leads to semidefinite programs that off-the-shelf interior-point method solvers can readily handle. For instance, each semidefinite program that was considered in the numerical applications of Section 6.4 was solved in under a second on a standard personal machine by the solver MOSEK [22]. An active and exciting area of research is focused on developing algorithms for large-scale semidefinite programs (see e.g. [138, 68]), and we believe that this effort can extend our learning approach to large-scale dynamical systems.

The size of our semidefinite programs is also affected by the degree of our candidate polynomial vector field and the degrees of the sos multipliers in (6.11) that result from the application of Putinar's Positivstellesatz. In practice, these degrees can be chosen using a statistical model validation technique, such as cross validation. These techniques take into account the fact that lower degrees can sometimes have a model regularization effect and lead to better generalization on unobserved parts of the state space.

We end by mentioning some questions that are left for future research.

- While the framework presented in this paper deals with continuous-time dynamical systems, we believe that most of the ideas could be extended to the discrete-time setting. It would be interesting to see how the definitions of side information, the approximation results, and the computational aspects contrast with the continuous-time case. Extending our framework to the problems of learning partial differential equations and stochastic differential equations with side information would also be interesting research directions.

- We have shown that for any $\delta > 0$, polynomial vector fields can approximate to arbitrary accuracy any vector field $f \in C_1^\circ(\Omega)$ while δ-satisfying any list of side information that f is known to satisfy. Even though from a practical standpoint, δ-satisfiability is sufficient (when δ is small), it is an interesting mathematical question in approximation theory to see which combinations of side information can be imposed exactly on polynomial vector fields while preserving an arbitrarily tight approximation guarantee to functions in $C_1^\circ(\Omega)$.

- We have presented a list of six types of side information that arise naturally in many applications and that lead to a convex formulation (meaning that a convex combination of two vector fields that satisfy any one of the six side information constraints will also satisfy the same side information constraint). There are of course other interesting side information constraints that do not lead to a convex formulation. Examples include the knowledge that an equilibirum point is locally or globally stable/stabilizable, and the knowledge that trajectories of the system starting in a set $A \subseteq \mathbb{R}^n$ avoid/reach another set $B \subseteq \mathbb{R}^n$. It is an interesting research direction to extend our approximation results and our sos-based approach to handle some of these nonconvex side information constraints.

- Finally, from a statistical and information-theoretic point of view, it is an interesting question to quantify the benefit of a particular side information constraint in reducing the number of trajectory observations needed to learn a good approximation of the unknown vector field.

Chapter 7

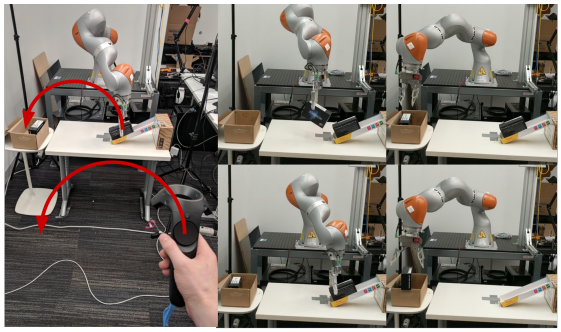
Teleoperator Imitation with Continuous-time Safety

7.1 Introduction

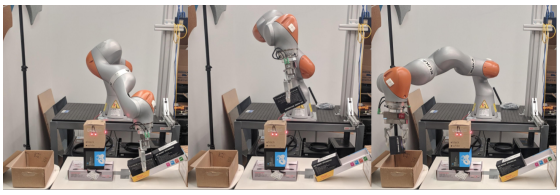
Teleoperation is enabling robotic systems to become pervasive in settings where full autonomy is currently out of reach [83, 210, 70]. Compelling applications include minimally invasive surgery [195, 197], space exploration [206], remote vehicle operations [79] and disaster relief scenarios [140]. A human teleoperator can control a robot through tasks that have complex semantics and are currently difficult to explicitly program or to learn to solve efficiently without supervision.

A downside of teleoperation is that it requires continuous error-free [26] operator attention even for highly repetitive tasks. This problem can be addressed through Learning-from-Demonstrations (LfD) or Imitation Learning techniques [182, 41] where a control law needs to be inferred from a small number of demonstrations. Such a law can then bootstrap data-efficient reinforcement learning for challenging tasks [210]. The demonstrator attempts to ensure that the robot's motions capture the relevant semantics of the task rather than requiring the robot to understand the semantics. The learnt control law should take over from the teleoperator and enable the robot to repeatedly execute the desired task even in dynamically changing conditions. For example, the origin of a picking task and the goal of a placing task may dynamically shift to configurations unseen during training, and moving obstacles may be encountered during execution. The latter is particularly relevant in collaborative human-robot workspaces where safety guarantees are paramount. In such situations, when faced with an obstacle, the robot cannot follow the demonstration path anymore and needs to recompute a new motion trajectory in real-time to avoid collision and still attempt to accomplish the desired task.

Work done as Google Internship.



(a) Pick and place teleoperation demonstration

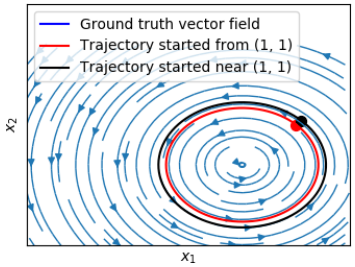


(b) Pick and place via contracting vector fields with obstacle avoidance.

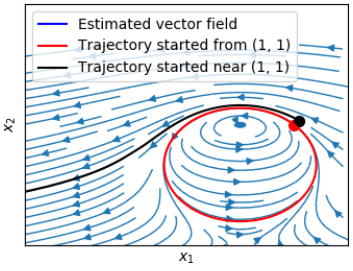
Figure 7.1: (a) A non-technical user provides a demonstrates via teleoperation to accomplish a pick and place task. (b) The robot now autonomously executes the pick and place task with a contracting vector field (CVF) allowing for continuous time guarantees while also avoiding obstacles.

Such real-time adaptivity can be elegantly achieved by associating demonstrations with a dynamical system [173, 189, 115, 113, 116]: a vector field defines a closed-loop velocity control law. From any state that the robot finds itself in, the vector field can then steer the robot back towards the desired imitation behavior, without the need for path replanning with classical approaches. Furthermore, the learnt vector field can be modulated in real-time [114, 100, 117] in order to avoid collisions with obstacles.

(a) Demo trajectory $\mathbf{x}^{\text{sample}}(t)$ started at (1,1)



(b) True Vector Field f^{real}



(c) Estimated vector field f^*

Figure 7.2: (a) a non-technical user demonstrates a circular trajectory. (b) the "ground truth" vector field. (c) the estimated vector field. Both vector fields produce the same trajectory when started from $(1,1)^T$ while they exhibit radically different behavior when started from a point arbitrarily close to $(1,1)^T$.

At first glance, the problem of imitation learning of a smooth dynamical system, $\dot{\mathbf{x}} = f(\mathbf{x})$ from samples $(\mathbf{x}, \dot{\mathbf{x}})$ appears to be a straightforward regression problem: simply minimize imitation loss $\sum_{i,t} \|f(\mathbf{x}^{(i)}(t)) - \dot{\mathbf{x}}^{(i)}(t)\|_2^2$ over a suitable family of vector valued maps, $f \in \mathcal{F}$. However, a naive supervised regression approach may be woefully inadequate, as illustrated in the middle panel of Figure 7.2 where the goal is to have a KUKA arm imitate a circular trajectory. As can be seen, estimating vector fields from a small number of trajectories potentially leads to instability – the

estimated field easily diverges when the initial conditions are even slightly different from those encountered during training. Therefore, unsurprisingly, learning with stability constraints has been the technical core of existing dynamical systems based LfD approaches, e.g. see [113, 116, 173, 189]. However, these methods have one or more of the following limitations: (1) they involve non-convex optimization for dynamics fitting and constructing Lyapunov stability certificates respectively and, hence, have no end-to-end optimality guarantees, (2) the notion of stability is not trajectory-centric, but rather focused on reaching a single desired equilibrium point, and (3) they are computationally infeasible when formulated in continuous-time. With this context, our contributions in this chapter include the following:

- We formulate a novel continuous time optimization problem over vector fields involving an imitation loss subject to a generalization-enforcing constraint that turns the neighborhood of demonstrations into contracting regions [133]. Within this region, all trajectories are guaranteed to coalesce towards the demonstration exponentially fast.

- We show that our formulation leads to an instance of time-varying semidefinite programming (see Chapter 2) for which a sum-of-squares relaxation [122, 157, 10] turns out to be exact! Hence, we can find the *globally optimal* polynomial vector field that has the lowest imitation loss among all polynomial vector fields of a given degree that are contracting on a region around the demonstrations in continuous time.

- On benchmark handwriting imitation tasks [113], our method outperforms competing approaches in terms of a variety of imitation quality metrics.

- We demonstrate our methods on a 7DOF KUKA pick-and-place LfD task where task completeness is accomplished despite dynamic obstacles in the environment, changing initial poses and moving destinations. By contrast, without contraction constraints, the vector field tends to move far from the demonstrated trajectory activating emergency breaks on the arm and failing to complete the task.

Our "dirty laundry" includes: (1) we cannot handle high degree polynomials due to the scalability limitations of current SDP solvers, and (2) our notion of incremental stability is local, even though our method generalizes well in the sense that a wide contraction tube is setup around the demonstrations.

7.2 Problem Statement

We are interested in estimating an unknown continuous time autonomous dynamical system

$$\dot{\mathbf{x}} = f^{\text{real}}(\mathbf{x}), \tag{7.1}$$

where $f^{\text{real}} : \mathbb{R}^n \to \mathbb{R}^n$ is an unknown continuously differentiable function.

We assume that we have access to one or several sample trajectories $\mathbf{x}^{(i)} : [0, T] \mapsto \mathbb{R}^n$ that satisfy $\dot{\mathbf{x}}^{(i)} = f^{\text{real}}(\mathbf{x}^{(i)})\ \forall t \in [0, T]$, where $T > 0$ and $i = 1, \ldots, M$. These

trajectories ($\mathbf{x}^{(i)}$, for $i = 1 \ldots, M$) constitute our *training data*, and our goal is to search for an approximation of the vector field f^{real} in a class of functions of interest $\mathcal{F}$ that reproduces trajectories as close as possible to the ones observed during training. In other words, we seek to solve the following continuous time least squares optimization problem (LSP):

$$f^* \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{M} \int_{t=0}^{T} \|f(\mathbf{x}^{(i)}(t)) - \dot{\mathbf{x}}^{(i)}(t)\|_2^2 \ \mathrm{d}t. \tag{LSP}$$

In addition to consistency with f^{real}, we want our learned vector field f to generalize in conditions that were not seen in the training data. Indeed, the LSP problem generally admits multiple solutions, as it only dictates how the vector field should behave on the sample trajectories. This under-specification can easily lead to overfitting, especially if the class of function $\mathcal{F}$ is expressive enough. The example of Figure 7.2 reinforces this phenomenon even for a simple circular motion. Note that standard data-independent regularization (e.g., L_2 regularizer) is insufficient to resolve the divergence illustrated here: a stronger stabilizer ensuring convergence, not just smoothness, of trajectories is needed. The notion of stability of interest to us in this chapter is *contraction* which we now briefly review.

7.2.1 Incremental stability and contraction analysis

Notions of stability called *incremental stability* and associated contraction analysis tools [104, 133] are concerned with the convergence of system trajectories with respect to each other, as opposed to classical Lyapunov stability which is with respect to a single equilibrium. Contraction analysis derives sufficient and necessary conditions under which the displacement between any two trajectories will go to zero. We give in this section a brief presentation of this notion based on [28].

Contraction analysis of a system $\dot{\mathbf{x}} = f(\mathbf{x})$ is best explained by considering the dynamics of $\delta\mathbf{x}(t)$, the infinitesimal displacement between two trajectories:

$$\delta\dot{\mathbf{x}} = \mathbf{J}_f(\mathbf{x})\delta\mathbf{x} \text{ where } \mathbf{J}_f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f.$$

From this equation we can easily derive the rate of change of the infinitesimal squared distance between two trajectories $\|\delta\mathbf{x}\|_2^2 = \delta\mathbf{x}^T\delta\mathbf{x}$ as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\delta\mathbf{x}\|_2^2 = 2\delta\mathbf{x}^T\delta\dot{\mathbf{x}} = 2\delta\mathbf{x}^T\mathbf{J}_f(\mathbf{x})\delta\mathbf{x}. \tag{7.2}$$

More generally, we can consider the infinitesimal squared distance with respect to a metric that is different from the Euclidian metric. A metric is given by smooth, matrix-valued function $\mathbf{M} : \mathbb{R}^+ \times \mathbb{R}^n \mapsto \mathbb{R}^{n\times n}$ that is uniformly positive definite, i.e. there exists $\varepsilon > 0$ such that

$$\mathbf{M}(t, \mathbf{x}) \succeq \varepsilon\mathbf{I} \quad \forall t \in \mathbb{R}^+, \ \forall \mathbf{x} \in \mathbb{R}^n, \tag{7.3}$$

where $\mathbf{I}$ is the identity matrix and the relation $A \succeq B$ between two symmetric matrices A and B is used to denote that the smallest eigenvalue of their difference $A - B$ is nonnegative. For the clarity of presentation, we only consider metric functions $\mathbf{M}(\mathbf{x})$ that do not depend on time t.

The squared norm of an infinitesimal displacement between two trajectories with respect to this metric is given by $\|\delta\mathbf{x}\|_{\mathbf{M}(\mathbf{x})}^2 := \delta\mathbf{x}^T\mathbf{M}(\mathbf{x})\delta\mathbf{x}$. The Euclidean metric corresponds to the case where $\mathbf{M}(\mathbf{x})$ is constant and equal to the identity matrix.

Similarly to (7.2), the rate of change of the squared norm of an infinitesimal displacement with respect to a metric $\mathbf{M}(\mathbf{x})$ follows the following dynamics:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\delta\mathbf{x}\|_{\mathbf{M}(\mathbf{x})}^2 = \delta\mathbf{x}^T(\mathbf{sym}[\mathbf{M}(\mathbf{x})\mathbf{J}_f(\mathbf{x})] + \dot{\mathbf{M}}(\mathbf{x}))\delta\mathbf{x}, \tag{7.4}$$

where $\mathbf{sym}[M]$ denotes $(M + M^T)/2$ for any square matrix M and $\dot{\mathbf{M}}(\mathbf{x})$ is the $n \times n$ matrix whose (i, j)-entry is $\nabla\mathbf{M}_{ij}(\mathbf{x})^T f(\mathbf{x})$. This motivates the following definition of *contraction*.

Definition 7.2.1 (Contraction)**.** *For a positive constant τ and a subset U of $\mathbb{R}^n$ the system $\dot{\mathbf{x}} = f(\mathbf{x})$ is said to be τ-contracting on the region U with respect to a metric $\mathbf{M}(\mathbf{x})$ if*

$$\mathbf{sym}[\mathbf{M}(\mathbf{x})\mathbf{J}_f(\mathbf{x})] + \dot{\mathbf{M}}(\mathbf{x}) \preceq -\tau\mathbf{M}(\mathbf{x}) \quad \forall\mathbf{x} \in U. \tag{7.5}$$

Remark 8. *When the vector field f is a linear function $\dot{\mathbf{x}} = A\mathbf{x}$, and the metric $\mathbf{M}(\mathbf{x})$ is constant $\mathbf{M}(\mathbf{x}) = P$, it is easy to see that contraction condition (7.5) is in fact equivalent to global stability condition,*

$$P \succ 0 \text{ and } \mathbf{sym}(PA^T) \preceq -\tau P. \tag{7.6}$$

Given a τ-*contracting* vector field with respect to a metric $\mathbf{M}(\mathbf{x})$, we can conclude from the dynamics in (7.4) that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\delta\mathbf{x}\|_{\mathbf{M}(\mathbf{x})}^2 \leq -\tau\|\delta\mathbf{x}\|_{\mathbf{M}(\mathbf{x})}$$

Integrating both sides yields,

$$\|\delta\mathbf{x}\|_{\mathbf{M}(\mathbf{x})} \leq e^{-\frac{\tau}{2}t}\|\delta\mathbf{x}(0)\|_{\mathbf{M}(\mathbf{x})}$$

Hence, any infinitesimal length $\|\delta\mathbf{x}\|_{\mathbf{M}(\mathbf{x})}$ (and by assumption (7.3), $\|\delta\mathbf{x}\|_2$) converges exponentially to zero as time goes to infinity. This implies that in a contraction region, trajectories will tend to converge together towards a nominal path. If the entire state-space is contracting and a finite equilibrium exists, then this equilibrium is unique and all trajectories converge to this equilibrium.

In the next section, we explain how to *globally* solve the following continuous-time vector field optimization problem to fit a contracting vector field to the training data given some fixed metric $\mathbf{M}(\mathbf{x})$. We refer to this as the least squares problem with contraction (LSPC):

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{M} \int_{t=0}^{T} \|f(\mathbf{x}^{(i)}(t)) - \dot{\mathbf{x}^{(i)}}(t)\|_2^2 \ \mathrm{d}t \tag{LSPC}$$
$$\text{s.t. } f \text{ is contracting on a region } U \subseteq \mathbb{R}^n$$
$$\text{containing the demonstrations } \mathbf{x}^{(i)}(t)$$
$$\text{with respect to the metric } \mathbf{M}(\mathbf{x}).$$

The search for a contraction metric itself may be interpreted as the search for a Lyapunov function of the specific form $V(\mathbf{x}) = f(\mathbf{x})^T \mathbf{M}(\mathbf{x}) f(\mathbf{x})$. As is the case with Lyapunov analysis in general, finding such an incremental stability certificate for a given dynamical system is a nontrivial problem; see [28] and references therein. If one wishes to find the vector field and a corresponding contraction metric at the same time, then the problem becomes non-convex. A common approach to handle this kind of problems is to optimize over one parameter at a time and fix the other one to its latest value and then alternate (i.e. fix a contraction metric and fit the vector field, then fix the vector field and improve on the contraction metric.)

7.3 Learning Contracting Vector Fields as a Time-Varying Convex Problem

In this section we explain how to formulate and solve the problem of learning a contracting vector field from demonstrations described in (LSPC). We will first see that we can formulate it as a *time-varying semidefinite problem*. We will then describe how to use tools from *sum of squares programming* to solve it.

7.3.1 Time-varying semidefinite problems

We call time-varying semidefinite problems (TV-SDP) optimization programs of the form

$$\begin{aligned} \min_{f \in \mathcal{F}} \quad & L(f) \\ \text{s.t.} \quad & \mathcal{L}_i f(t) \succeq 0 \quad \forall i = 1, \dots, m \ \forall t \in [0, T], \end{aligned} \tag{TV-SDP}$$

where the variable $t \in [0, T]$ stands for time, the loss function $L : \mathcal{F} \mapsto \mathbb{R}$ in the objective is assumed to be convex and the $\mathcal{L}_i$ $(i = 1, \dots, m)$ are linear functionals that map an element $f \in \mathcal{F}$ to a matrix-valued function $\mathcal{L}_i f : [0, T] \mapsto \mathbb{R}^{n \times n}$. We will restrict the space of functions $\mathcal{F}$ to be the space of functions whose components are polynomials of degree $d \in \mathbb{N}$:

$$\mathcal{F} := \{f : \mathbb{R}^n \mapsto \mathbb{R}^n \mid f_i \in \mathbb{R}_d[\mathbf{x}]\}, \tag{7.7}$$

and we make the assumption that $\mathcal{L}_i f$ is a matrix with polynomial entries. Our interest in this setting stems from the fact that polynomial functions can approximate most functions reasonably well. Moreover, polynomials are suitable for algorithmic operations as we will see in the next section. See [5] for a more in-depth treatment of time-varying semidefinite programs with polynomial data.

Let us now show how to reformulate the problem in (LSPC) of fitting a vector field $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ to m sample trajectories $\{(\mathbf{x}^{(i)}(t), \dot{\mathbf{x}}^{(i)}(t)) \mid t \in [0, T],\ i = 1, \ldots, m\}$ as a (TV-SDP). For this problem to fit within our framework, we start by approximating each trajectory $\mathbf{x}^{(i)}(t)$ with a polynomial function of time $\mathbf{x}^{(i)}_{\text{poly}}(t)$. Our decision variable is the polynomial vector field f and we seek to optimize the following objective function

$$L(f) := \sum_{i=1}^{M} \int_{t=0}^{T} \|f(\mathbf{x}^{(i)}_{\text{poly}}(t)) - \dot{\mathbf{x}}^{(i)}_{\text{poly}}(t)\|_2^2 \, \mathrm{d}t \tag{7.8}$$

which is already convex (in fact convex quadratic). In order to impose the contraction of the vector field f over some region around the trajectories in demonstration, we use a smoothness argument to claim that it is sufficient to impose contraction *only on* the trajectories themselves. See Proposition 7.3.4 later for a more quantitative statement of this claim. To be concrete, we take

$$\begin{aligned} \mathcal{L}_i f(\cdot) := &- \mathbf{sym}[\mathbf{M}(\mathbf{x}^{(i)}_{\text{poly}}(\cdot))\mathbf{J}_f(\mathbf{x}^{(i)}_{\text{poly}}(\cdot))] \\ &- \dot{\mathbf{M}}(\mathbf{x}^{(i)}_{\text{poly}}(\cdot)) - \tau \mathbf{M}(\mathbf{x}^{(i)}_{\text{poly}}(\cdot)), \end{aligned} \tag{7.9}$$

where $\mathbf{M}(\mathbf{x})$ is some known contraction metric.

7.3.2 Sum-of-squares programming

In this section we review the notions of sum-of-squares (SOS) programming and its applications to polynomial optimization, and how we apply it for learning a contracting polynomial vector field. SOS techniques have found several applications in Robotics: constructing Lyapunov functions [2], locomotion planning [164], design and verification of provably safe controllers [137], grasping and manipulation [65], inverse optimal control [159] and modeling 3D geometry [13].

Let $\mathbb{R}_d[\mathbf{x}]$ be the ring of polynomials $p(\mathbf{x})$ in real variables $\mathbf{x} = (x_1, \ldots, x_n)$ with real coefficients of degree at most d. A polynomial $p \in \mathbb{R}[\mathbf{x}]$ is nonnegative if $p(\mathbf{x}) \geq 0$ for every $\mathbf{x} \in \mathbb{R}^n$. In many applications, including the one we cover in this chapter, we seek to find the coefficients of one (or several) polynomials without violating some nonnegativity constraints. While the notion of nonnegativity is conceptually easy to understand, even testing whether a given polynomial is nonnegative is known to be NP-hard as soon as the degree $d \geq 4$ and the number of variables $n \geq 3$.

A polynomial $p \in \mathbb{R}_d[\mathbf{x}]$, with d even, is a sum-of-squares (SOS) if there exists polynomials $q_1, \ldots, q_m \in \mathbb{R}_{\frac{d}{2}}[\mathbf{x}]$ such that

$$p(\mathbf{x}) = \sum_{i=1}^{m} q_i(\mathbf{x})^2. \tag{7.10}$$

An attractive feature of the set of SOS polynomials is that optimizing over it can be cast as a semidefinite program of tractable size, for which many solvers already exist. Indeed, it is known [122][157] that a polynomial $p(\mathbf{x})$ of degree d can be decomposed as in (7.10) if and only if there exists a positive semidefinite matrix Q such that

$$p(\mathbf{x}) = z(\mathbf{x})^T Q z(\mathbf{x}) \ \forall \mathbf{x} \in \mathbb{R}^n,$$

where $z(\mathbf{x})$ is the vector of monomials of $\mathbf{x}$ up to degree $\frac{d}{2}$, and the equality between the two sides of the equation is equivalent to a set of linear equalities in the coefficients of the polynomial $p(\mathbf{x})$ and the entries of the matrix Q.

Sum-of Squares Matrices: If a polynomial $p(\mathbf{x})$ is SOS, then it is obviously nonnegative, and the matrix Q acts as a certificate of this fact, making it easy to check that the polynomial at hand is nonnegative for every value of the vector $\mathbf{x}$. In order to use similar techniques to impose contraction of a vector field, we need a slight generalization of this concept to ensure that a *matrix-valued* polynomial $P(\mathbf{x})$ (i.e. a matrix whose entries are polynomial functions) is positive semidefinite (PSD) for all values of $\mathbf{x}$. We can equivalently consider the scalar-valued polynomial $p(\mathbf{x}, \mathbf{u}) := \mathbf{u}^T P(\mathbf{x}) \mathbf{u}$, where $\mathbf{u}$ is a $n \times 1$ vector of new indeterminates, as positive semidefiniteness of $P(\mathbf{x})$ is equivalent to the nonnegativity of $p(\mathbf{x}, \mathbf{u})$. If $p(\mathbf{x}, \mathbf{u})$ is SOS, then we say that P is *a sum-of-squares matrix* (SOSM) [118, 81, 186]. Consequently, optimizing over SOSM matrices is a tractable problem.

Exact Relaxation: A natural question here is how much we lose by restricting ourselves to the set of SOSM matrices as opposed the set of PSD matrices. In general, these two sets are quite different [54]. In our case however, all the matrices considered are univariate as they depend only on the variable of time t. *It turns out that, in this special case, these two notions are equivalent!*

Theorem 7.3.1 (See e.g. [55]). *A matrix-valued polynomial $P(t)$ is PSD everywhere (i.e. $P(t) \succeq 0 \ \forall t \in \mathbb{R}$) if and only if the associated polynomial $p(t, \mathbf{u}) := \mathbf{u}^T P(t) \mathbf{u}$ is SOS.*

The next theorem generalizes this result to the case where we need to impose PSD-ness only on the interval $[0, T]$ (as opposed to $t \in \mathbb{R}$.)

Theorem 7.3.2 (See Theorem 2.5 of [67]). *A matrix-valued polynomial $P(t)$ of degree d is PSD on the interval $[0, T]$ (i.e. $P(t) \succeq 0 \ \forall t \in [0, T]$) if and only if can be written as*

$$\begin{cases} P(t) &= tV(t) + (T-t)W(t) \quad \textit{if } \deg(P) \textit{ odd,} \\ P(t) &= V(t) + t(T-t)W(t) \quad \textit{if } \deg(P) \textit{ even.} \end{cases}$$

where $V(t)$ and $W(t)$ are SOSM. In the first case, $V(t)$ and $W(t)$ have degree at most $\deg(P)-1$, and in the second case $V(t)$ (resp. $W(t)$) has degree at most $\deg(P)$ (resp. $\deg(P)-2$). When that is the case, we say that $P(t)$ is SOSM on $[0,T]$.

7.3.3 Main result and CVF-P

The main result of this section is summarized in the following theorem that states that the problem of fitting a contracting polynomial vector field to polynomial data can be cast as a semidefinite program.

Theorem 7.3.3. *The following semidefinite program*

$$\begin{aligned} \min_{f\in\mathcal{F}} \quad & \sum_{i=1}^{M}\int_{t=0}^{T} \|f(\mathbf{x}_p^{(i)}(t)) - \dot{\mathbf{x}}_p{}^{(i)}(t)\|_2^2 \ \mathrm{d}t \\ s.t. \quad & \mathcal{L}_i f \text{ is SOSM on } [0,T] \text{ for } i=1,\dots,M. \end{aligned} \tag{LSPC-SOS}$$

with $\mathcal{F}$, $\mathcal{L}_i$, and L defined as in (7.7), (7.9) and (7.8) resp. finds the polynomial vector field that has the lowest fitting error $L(f)$ among all polynomial vector fields of degree d that are contracting on a region containing the demonstrations $\mathbf{x}_p^{(i)}$.

To reiterate, the above sum-of-squares relaxation leads to no loss of optimality: the SDP above returns the globally optimal solution to the problem stated in LSPC. Our numerical implementation uses the Splitting Conic Solver (SCS) [147] for solving large-scale convex cone problems.

Remark 9. *Note that the time complexity of solving the SDP defined in (LSPC-SOS) is bounded above by a polynomial function of the number of trajectories, the dimension n of the space where they live, and the degree d of the candidate polynomial vector field. In practice however, only small to moderate values for n and d can be solved for as the exponents appearing in this polynomial are prohibitively large. Significant progress has been made in recent years in inventing more scalable alternatives to SDPs based on linear and second order cone programming that can be readily applied to our framework [8].*

For the rest of this chapter, our approach will be abbreviated as CVF-P, standing for Polynomial Contracting Vector Fields.

7.3.4 Generalization properties

The contraction property of CVF-P generalizes to a wider region in the state space. The next proposition shows that any sufficiently smooth vector field that is feasible for the problem stated in LSPC-SOS is contracting on a "tube" around the demonstrated trajectories.

Proposition 7.3.4 (A lower bound on the contraction tube)**.** *If $f:\Omega\subseteq\mathbb{R}^n\mapsto\mathbb{R}^n$ is a twice continuously differentiable vector field that satisfies*

$$-\mathbf{sym}[\mathbf{M}(\mathbf{x}(t))\mathbf{J}_f(\mathbf{x}(t))] - \dot{\mathbf{M}}(\mathbf{x}(t)) \succeq \tau\mathbf{M}(\mathbf{x}) \quad \forall t\in[0,T]$$

where Ω is a compact region of $\mathbb{R}^n$, τ is a positive constant, $\mathbf{M}(\mathbf{x})$ is a positive definite metric, and $\mathbf{x} : [0,T] \mapsto \mathbb{R}^n$ is a path, then f is $\tau/2$-contracting with respect to the metric $\mathbf{M}(\mathbf{x})$ on the region U defined by

$$U := \{\mathbf{x}(t) + \delta \mid t \in [0,T],\ \|\delta\|_2 \leq \varepsilon\} \cap \Omega,$$

where ε is positive scalar depending only τ and on the smoothness parameters of $f(\mathbf{x})$ and $\mathbf{M}(\mathbf{x})$ and is defined explicitly in Eqn. 7.11.

For the proof we will need the following simple fact about symmetric matrices.

Lemma 7.3.5. *For any $n \times n$ symmetric matrices A and B*

$$|\lambda_{\min}(A) - \lambda_{\min}(B)| \leq n \max_{ij} |A_{ij} - B_{ij}|,$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue function.

Proof of Proposition 7.3.4. Let f, $\mathbf{M}$, Ω and τ be as in the statement of Proposition 7.3.4. Define $c := \min_{\mathbf{x}\in\Omega} \lambda_{\min}(\mathbf{M}(\mathbf{x}))$. Notice that since the metric $\mathbf{M}(\mathbf{x})$ is uniformly positive definite, then $c > 0$. Let us now define

$$\varepsilon := \frac{\tau c}{2nK} > 0 \tag{7.11}$$

where K is the scalar equal to

$$\max_{1\leq i,j\leq n} \sup_{\mathbf{x}\in\Omega} \| \frac{\partial}{\partial \mathbf{x}} \left(\mathbf{sym}[\mathbf{M}(\mathbf{x})\mathbf{J}_f(\mathbf{x})] + \dot{\mathbf{M}}(\mathbf{x}) - \frac{\tau}{2}\mathbf{M}(\mathbf{x}) \right)_{ij} \|_2.$$

Fix $t \in [0,T]$, and let δ be a vector in $\mathbb{R}^n$ such that $\|\delta\|_2 \leq \varepsilon$. Our aim is to prove that the matrix R^δ defined by

$$-\mathbf{sym}[\mathbf{M}(\mathbf{x}(t)+\delta)\mathbf{J}_f(\mathbf{x}(t)+\delta)] - \dot{\mathbf{M}}(\mathbf{x}(t)+\delta) - \frac{\tau}{2}\mathbf{M}(\mathbf{x}(t)+\delta)$$

is positive semidefinite. Notice that our choice for K guarantees that the maps $\delta \mapsto R^\delta_{ij}$ are L-Lipchitz for $i,j = 1,\ldots,n$, therefore $\max_{ij} |R^\delta_{ij} - R^0_{ij}| \leq K\varepsilon$. Using Lemma 7.3.5 we conclude that the smallest eigenvalues of R^δ and R^0 are within a distance of $nK\varepsilon$ of each other. Since we assumed that $R^0 \succeq \frac{\tau}{2}\mathbf{M}(\mathbf{x}(t))$, then $\lambda_{\min}(R^0)$ is at least $c\frac{\tau}{2}$, and therefore $\lambda_{\min}(R^\delta)$ is at least $c\frac{\tau}{2} - nK\varepsilon$. We conclude that our choice of ε in (7.11) guarantees that R^δ is positive semidefinite. □

We note that the estimate obtained in this proposition is quite conservative. In practice the contraction tube is much larger than what is predicted here.

7.4 Empirical Comparisons: Handwriting Imitation

We evaluate our methods on the LASA library of two-dimensional human handwriting motions commonly used for benchmarking dynamical systems based movement generation techniques in imitation learning settings [116][128][173]. This dataset contains 30 handwriting motions recorded with a pen input on a Tablet PC. For each motion, the user was asked to draw 7 demonstrations of a desired pattern, by starting from different initial positions and ending at the same final point. Each demonstration trajectory comprises of 1000 position ($\mathbf{x}$) and velocity ($\dot{\mathbf{x}}$) measurements. We use 4 demonstrations for training and 3 demonstrations for testing as shown in Figure 7.3.

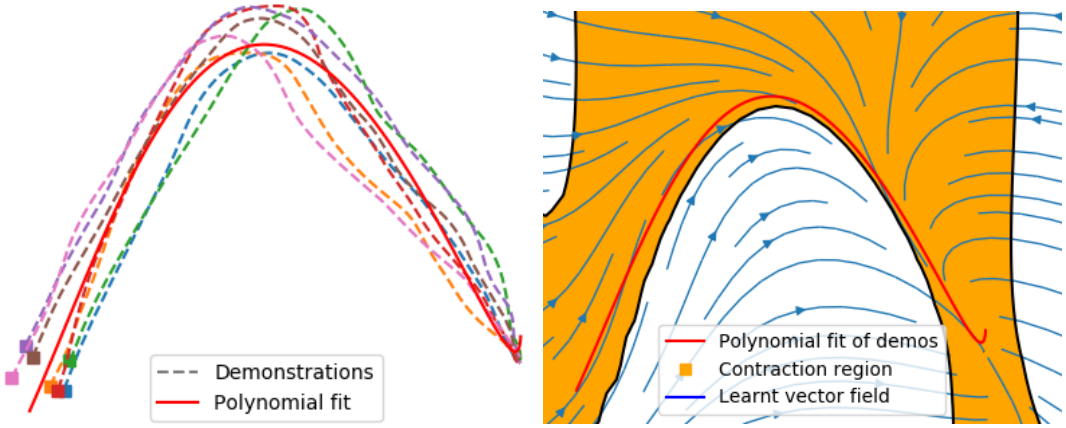


Figure 7.3: The figure on the left shows demonstration trajectories (dotted) and the polynomial fit of the demonstrations (solid line) for the *Angle* shape. The figure on the right visualizes both the polynomial fit (red), the learnt vector field (blue), and the contraction region (orange) for the incrementally stable vector field learned using our method.

We report in Table 7.1 comparisons on the *Angle* shape against state of the art methods for estimating stable dynamical systems, the Stable Estimator of Dynamical Systems (SEDS) [113], Control Lyapunov Function-based Dynamic Movements (CLFDM) [115] and Dynamic Movement Primitives (DMP) [101]. The training process in these methods involves non-convex optimization with no global optimality guarantees. Additionally, DMPs can only be trained from one demonstration one degree-of-freedom at a time. For all experiments, we learn degree 5 CVF-Ps with $\tau = 1.0$ and $\mathbf{M}(\mathbf{x}) = \mathbf{I}$. We report the following imitation quality metrics.

Reproduction Accuracy: How well does the vector field reproduce positions and velocities in training and test demonstrations, when started from same initial conditions and integrated for the same amount of time as the human movement duration (T). Specifically, we measure reproduction error with respect to m demonstration

Metric	DMP	SEDS	CLFDM	**CVF-P**
Reproduction Accuracy				
TrainingTrajectoryError	4.1	7.2	4.9	6.5
TrainingVelocityError	7.4	14.6	11.0	13.9
TestTrajectoryError	5.5	4.6	12.2	3.8
TestVelocityError	8.7	11.3	15.5	11.4
Stability				
DistanceToGoal	3.6	3.2	6.7	2.5
DurationToGoal	-	3.9	4.3	3.3
NumberReachedGoal	0/7	7/7	7/7	7/7
GridDuration (sec)	5.9	3.7	9.7	1.9
GridFractionReachedGoal	6%	100%	100%	100%
GridDistanceToGoal	3.3	1.0	1.0	1.0
GridDTWD ($\times 10^4$)	2.4	1.4	1.4	2.0
Training and Integration Speed (in seconds)				
TrainingTime	0.05	2.1	2.8	0.2
IntegrationSpeed	0.21	0.06	0.15	0.01

Table 7.1: LASA Angle shape benchmarks. Our approach is CVF-P.

trajectories as,

$$\text{TrajectoryError} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T_i}\sum_{t=0}^{T_i}\|\mathbf{x}^i(t) - \hat{\mathbf{x}}^i(t)\|_2$$

$$\text{VelocityError} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T_i}\sum_{t=0}^{T_i}\|\dot{\mathbf{x}}^i(t) - \hat{\dot{\mathbf{x}}}^i(t)\|_2.$$

The metrics *TrainingTrajectoryError, TestTrajectoryError, TrainingVelocityError, TestVelocityError* report these measures with respect to training and test demonstrations. At the end of the integration duration (T), we also report *DistanceToGoal*: how far the final state is from the goal (origin). Finally, to account for the situation where the learnt dynamics is somewhat slower than the human demonstration, we also generate trajectories for a much longer time horizon ($30T$) and report *DurationToGoal*: the time it took for the state to enter a ball of radius $1mm$ around the goal, and how often this happened for the 7 demonstrations (*NumReachedGoal*).

Stability: To measure stability properties, we evolve the dynamical system from 16 random positions on a grid enclosing the demonstrations for a long integration time horizon (30T). We report the fraction of trajectories that reach the goal ($GridFraction$); the mean duration to reach the goal when that happens ($GridDuration$); the mean distance to the Goal ($GridDistanceToGoal$) and the closest proximity of the generated trajectories to a human demonstration, as measured using Dynamic Time Warping Distance ($GridDTWD$) [111] (since in this case trajectories are likely of lengths different from demonstrations).

Training and Integration Speed: We measure both training time as well as time to evaluate the dynamical system which translates to integration speed.

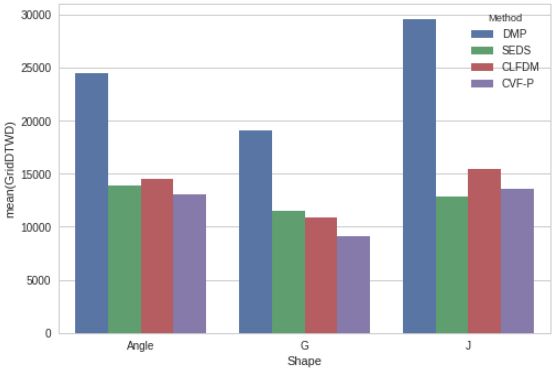


Figure 7.4: GridDTWD comparison on Angle, G and J shapes.

(a) Home (b) Pick (c) Place

Figure 7.5: In our task, the robot must move between the (a) home to (b) pick, (c) a place positions.

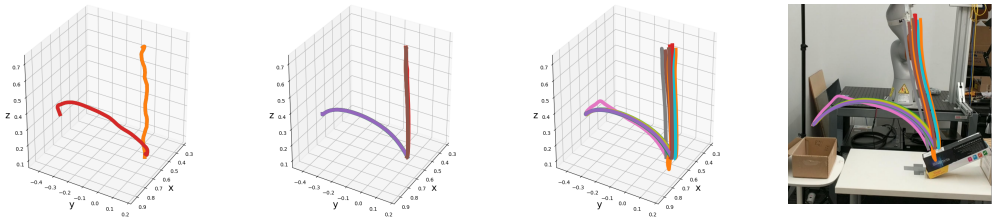
It can be seen that our approach is highly competitive on most metrics: reproduction quality, stability, and training and inference speed. In particular, it returns the best mean dynamic time warping distance to the demonstrations when initialized from points on a grid. A comparison of *GridDTWD* on a few other shapes is shown in Figure. 7.4.

7.5 Pick-and-Place with Obstacles

We consider a kitting task shown in Figure 7.5 where objects are picked from a known location and placed into a box. A teleoperator quickly demonstrates a few trajectories guiding a 7DOF KUKA IIWA arm to grasp objects and place them in a box. After learning from demonstrations, the robot is expected to continually fill boxes to be verified and moved by a human working in close proximity freely moving obstacles in and out of the workspace. The arm is velocity-controlled in joint space at 50 Hz.

7.5.1 Demonstration trajectory

Figure 7.6a shows the demonstration pick and place trajectory collected from the user. This trajectory was collected using an HTC Vive controller operated by a user standing in front of and watching the robot move through the demonstration as it is produced. Different buttons on the remote were used to open/close the gripper, send

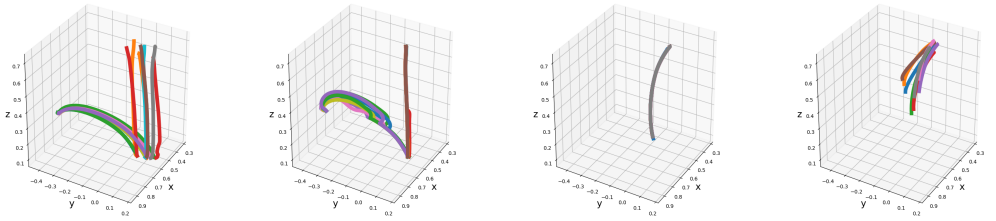


(a) Demonstration Trajectory

(b) CVF-P

(c) CVF-P, 0.05 Noise

(d) CVF-P, 0.05 Noise



(e) CVF-P, 0.1 Noise

(f) CVF-P and Obstacle

(g) VF-P, No Contraction, No Noise, No Obstacle

(h) VF-P, No Contraction, 0.05 Noise, No Obstacle

Figure 7.6: (a) A user demonstrated trajectory visualization shows the path of the end effector through cartesian space. (b) Eight trajectories executed using a vectorfield in joint space learned from the demonstration. (c,d) Eight trajectories with uniform noise between [-0.05, 0.05] radians was added per-joint to the initial joint state. (e) Eight trajectories with uniform noise between [-0.1, 0.1] added to the initial joint state. (f) Eight new trajectories with an object in the way that modulates the learned vector field. Notice the motion deviates, and then returns to the desired trajectory. (g) Eight trajectories without contraction, the arm deviates from the demonstration and cannot complete the trajectory. (h) Eight trajectories without contraction and [-0.05, 0.05] noise, the arm cannot complete the trajectory.

the arm to the Home position, and indicate the start of a new trajectory. The pick and place task was collected as two separate trajectories, one for the pick motion and another for the place motion.

7.5.2 Learning a composition of pick and place CVF-Ps

Using the demonstration trajectory, two different polynomial contracting vector fields (CVF-Ps) were fit to the data, one for the pick motion, one for the place. These trajectories were fit to a degree 2 polynomial with $\tau = 0.1$ and $\mathbf{M}(\mathbf{x}) = \mathbf{I}$, using an SCS solver run for 2500 iterations. For the ease of visualization, we show the trajectories in cartesian space in Figure 7.6. The CVF-P was fit to the trajectory in the 7-dimensional joint space. The arm was then run through using the vector field eight times starting from the home position. Each trajectory was allowed to run until the L_2-norm of the arm joint velocities dropped below a threshold of 0.01. At that point, the arm would begin to move using the second vector field. The trajectories

taken by the arm are shown in Figure 7.6b. The eight runs have very little deviation from each other.

7.5.3 Generalization to different initial poses

Next, noise is added to the home position of the arm, and again the vector field is used to move the arm through the task. Figure 7.6c noise is added uniformly from the range [-0.05, 0.05] radians to each value of each joint of the arm's starting home position. Figure 7.6d, shows these same trajectories overlaid on the Kuka arm. In Figure 7.6e uniform noise is added in the same manner from the range $[-0.1, 0.1]$. Due to contraction, trajectories are seen to converge from random initial conditions.

7.5.4 What happens without contraction constraints?

In Figure 7.6g the arm is run eight times using a vector field without contraction. While the arm is consistent in the trajectory that it takes, the arm moves far from the demonstrated trajectory, and eventually causes the emergency break to activate at joint limits, failing to finish the task.

In Figure 7.6h The arm is again run eight times without contraction with noise added uniformly from the range [-0.05, 0.05] to each the value of each joint of the arm's starting home position. The trajectory of the arm varies widely and had to be cut short as it was continually causing the emergency break to engage.

7.5.5 Whole-body obstacle avoidance

Here we enable a Kuka robot arm to follow demonstrated trajectories while avoiding obstacles unseen during training. In the system we describe below, collisions are avoided against any part of robot body. At every timestep, a commodity depth sensor like the Intel RealSense or PhaseSpace motion capture acquires a point cloud representation of the obstacle. Our setup is along the lines of [109], although we do not model occluded regions as occupied. At this point, our demonstrations and trajectories exist in joint space $\mathcal{J} \approx \mathbb{R}^7$, while our obstacle pointclouds exists in Cartesian space $\mathcal{C} \approx \mathbb{R}^3$ with an origin at the base of the robot.

Cartesian to joint space Map

We pre-compute a set-valued *inverse kinematic map* IK that maps a position $c \in \mathcal{C}$ to a subset of $\mathcal{J}$ containing all the joint configurations that would cause any part of the arm to occupy the position c.

More formally, the obstacles positions are known in Cartesian space $\mathcal{C}$ different from the control space $\mathcal{J}$ of the robot. (e.g. we control the joint angles rather than end-effector pose.) The Kuka arm simulator allows us to query the forward kinematics map $\text{FK} : \mathcal{J} \to \mathcal{C}$. To compute the inverse of this map, the joint space of the robot was discretized into 658,945 discrete positions. These discrete positions were created by

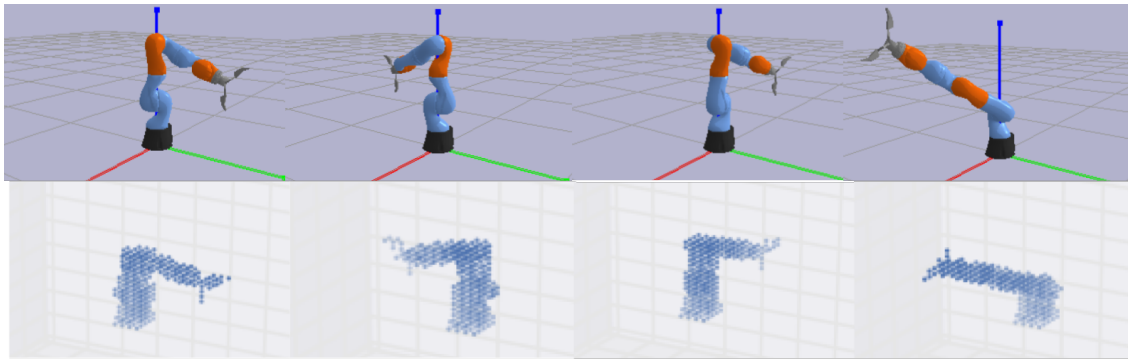


Figure 7.7: In order to produce a cartesian to joint space mapping, pybullet[62] was used to place the arm in over 658,945 configurations such as the 4 in the top row. Then a voxelization of the arm was produced in this pose using binvox.

regularly sampling each joint from a min to max angle using a step size of 0.1 radians. As shown in Figure 7.7, the robot was positioned at each point of the 658,945 discrete joint space points within pybullet[62], and the robot was voxelized using binvox[141]. This produced the map FK. We then compute IK := FK^{-1}.

Modulation of contracting vector fields

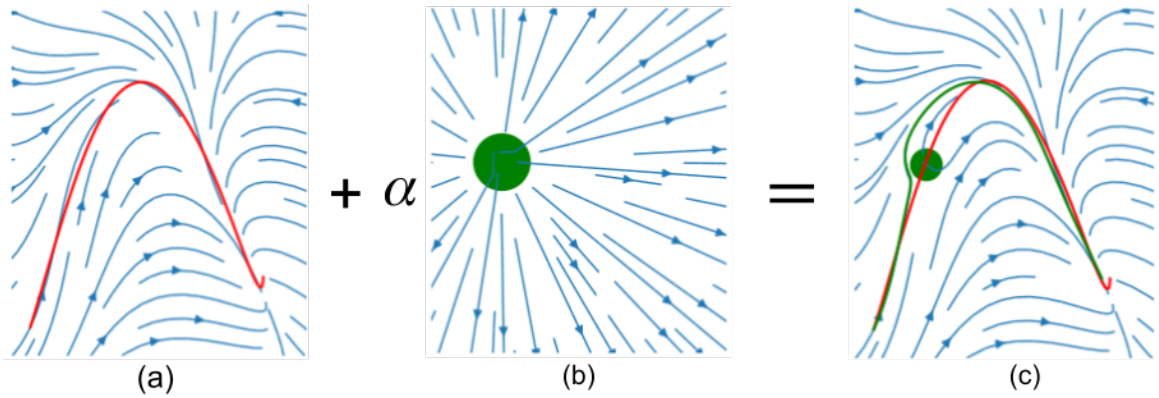


Figure 7.8: (a) Shows a vector field f learnt from a nominal path (red). (b) Depicts a repulsive vector field $h^{\text{obstacles}}$ associated with an obstacle (green disk). (c) Shows modulated vector field $\tilde{f}$ (blue) plotted with a sample trajectory (green).

The obstacle positions are then incorporated in a repulsive vector-field to push the arm away from collision as it moves,

$$h^{\text{obstacles}}(t, \mathbf{x}) := \sum_{\substack{\text{positions of} \\ \text{obstacles } c \\ \text{at time } t}} \sum_{j \in T^{-1}(c)} \frac{\mathbf{x} - j}{\|\mathbf{x} - j\|_2^r}, \tag{7.12}$$

where the integer r control how fast the effect of this vector field decays as a function of distance (a high value of r makes the effect of $h^{\text{obstacles}}$ local, while a small value

makes its effect more uniform.) This vector field is added to our learnt vector-field f to obtain a *modulated vector field* (depicted in Figure 7.8)

$$\tilde{f}(t, \mathbf{x}) = f(\mathbf{x}) + \alpha \, h^{\text{obstacles}}(t, \mathbf{x}),$$

where α is positive constant that is responsible for controlling the strength of the modulation, that is then fed to the Kuka arm. If the modulation is local and the obstacle is well within the joint-space contraction tube, we expect the motion to re-converge to the demonstrated behavior.

We point out that it is possible to use alternative modulation methods that come with different guarantees and drawbacks. In [114, 100] for instance, the authors use a multiplicative modulation function that preserves equilibrium points in the case of convex or concave obstacles.

While our approach does not enjoy the same guarantees, its additive nature allows us to handle a large number of obstacles as every term in Eqn. 7.12 can be computed in a distributed fashion, and furthermore, we do not need to impose any restrictions on the shape of the obstacles (convex/concave). This is particularly important as our control space $\mathcal{J}$ is different from the space $\mathcal{C}$ where the obstacle are observed, and the map IK that links between the two spaces can significantly alter the shape of an obstacle in general (e.g. a sphere in cartesian space can be mapped to a disconnected set in joint space).

Real-time obstacle avoidance

Here, using a *real-time motion capture system*, an obstacle is introduced to the robot's workspace as shown in Figure 7.1b. Eight trajectories were executed from the home position with the obstacle in the workspace, and the resultant trajectories are shown in Figure 7.6f. At each timestep, the objects position was returned by the motion capture system. The point in Cartesian space was used to modulate the joint space vectorfield as described in Section 7.5.5. The tasks are accomplished as the arm avoids obstacles but remains within the joint-space contraction tube re-converging to the demonstrated behavior.

7.6 Conclusion

This work presents a novel approach to teleoperator imitation using contracting vector fields that are globally optimal with respect to loss minimization and providing continuous-time guarantees on the behaviour of the system when started from within a contraction tube around the demonstration. Our approach compares favorably with other movement generation techniques. Additionally, we build a workspace cartesian to joint space map for the robot, and utilize it to update our CVF on the fly to avoid dynamic obstacles. We demonstrate how this approach enables the transfer of knowledge from humans to robots for accomplishing a real world robotic pick and place task. Future work includes greater scalability of our solution, composition of CVFs

for more complex tasks, integrating with a perception module and helping bootstrap data-hungry reinforcement learning approaches.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*. Elsevier, 2003.

[2] A. A. Ahmadi. *Algebraic relaxations and hardness results in polynomial optimization and Lyapunov analysis*. PhD thesis, MIT, 2011.

[3] A. A. Ahmadi. On the difficulty of deciding asymptotic stability of cubic homogeneous vector fields. In *Proceedings of the American Control Conference*, pages 3334–3339, 2012.

[4] A. A. Ahmadi and B. El Khadir. On algebraic proofs of stability for homogeneous vector fields. *IEEE Transactions on Automatic Control 65 (1), 325-332*, 2019.

[5] A. A. Ahmadi and B. El Khadir. Time-varying semidefinite programs. *Preprint available at arXiv:1808.03994. To appear in Mathematics of Operations Research*, 2020.

[6] A. A. Ahmadi and B. El Khadir. Learning dynamical systems with side information (short version). In *Proceedings of the* 2^{nd} *Conference on Learning for Dynamics and Control*, volume 120, pages 718–727, 2020.

[7] A. A. Ahmadi and A. Majumdar. Some applications of polynomial optimization in operations research and real-time decision making. *Optimization Letters*, 10 (4):709–729, 2016.

[8] A. A. Ahmadi and A. Majumdar. DSOS and SDSOS optimization: more tractable alternatives to sum of squares and semidefinite optimization. *SIAM Journal on Applied Algebraic Geometry*, 3(193), 2019.

[9] A. A. Ahmadi and P. A. Parrilo. A complete characterization of the gap between convexity and sos-convexity. *SIAM Journal on Optimization*, 23(2):811–833, 2013.

[10] A. A. Ahmadi and P. A. Parrilo. Towards scalable algorithms with formal guarantees for lyapunov analysis of control systems via algebraic optimization. In *2014 IEEE 53rd Annual Conference on Decision and Control (CDC)*, pages 2272–2281. IEEE, 2014.

[11] A. A. Ahmadi and P. A. Parrilo. Sum of squares certificates for stability of planar, homogeneous, and switched systems. *IEEE Transactions on Automatic Control*, 62(10):5269–5274, 2017.

[12] A. A. Ahmadi, M. Krstic, and P. A. Parrilo. A globally asymptotically stable polynomial vector field with no polynomial Lyapunov function. In *Proceedings of the IEEE Conference on Decision and Control*, pages 7579–7580, 2011.

[13] A. A. Ahmadi, G. Hall, A. Makadia, and V. Sindhwani. Geometry of 3d environments and sum of squares polynomials. 2016.

[14] M. Ahmadi, U. Topcu, and C. Rowley. Control-oriented learning of Lagrangian and Hamiltonian systems. In *Annual American Control Conference*, pages 520–525, 2018.

[15] H. Anai and P. A. Parrilo. Convex quantifier elimination for semidefinite programming. In *Proceedings of the International Workshop on Computer Algebra in Scientific Computing, CASC*, 2003.

[16] E. J. Anderson and P. Nash. *Linear programming in infinite-dimensional spaces: theory and applications*. John Wiley & Sons, 1987.

[17] E. J. Anderson and A. B. Philpott. A continuous-time network simplex algorithm. *Networks*, 19(4):395–425, 1989.

[18] R. M. Anderson, B. Anderson, and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[19] A. Andreini, A. Bacciotti, and G. Stefani. Global stabilizability of homogeneous vector fields of odd degree. *Systems and Control Letters*, 10(4):251–256, 1988.

[20] M. F. Anjos and J. B. Lasserre. *Handbook on semidefinite, conic and polynomial optimization*, volume 166. Springer Science & Business Media, 2011.

[21] K. M. Anstreicher. Generation of feasible descent directions in continuous time linear programming. *Tech. Report, SOL 83-18, Department of Operations Research, Stanford University, Stanford, CA*, 1984.

[22] M. ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 8.1.* 2017.

[23] J. Argémi. Sur les points singuliers multiples de systèmes dynamiques dans R2. *Ann. Mat. Pure Appl.*, 79(1):35–69, 1968.

[24] V. I. Arnold. Algebraic unsolvability of the problem of Lyapunov stability and the problem of topological classification of singular points of an analytic system of differential equations. *Functional Analysis and its Applications*, 4(3):173–180.

[25] V. I. Arnold. Problems of present day mathematics, XVII (Dynamical systems and differential equations). *Proc. Symp. Pure Math.*, 28(59), 1976.

[26] C. G. Atkeson, B. Babu, N. Banerjee, D. Berenson, C. Bove, X. Cui, M. DeDonato, R. Du, S. Feng, P. Franklin, et al. What happened at the DARPA robotics challenge, and why. 2015.

[27] E. M. Aylward, S. M. Itani, and P. A. Parrilo. Explicit SOS decompositions of univariate polynomial matrices and the Kalman-Yakubovich-Popov lemma. In *Proceedings of the 46th IEEE Conference on Decision and Control*, pages 5660–5665, 2007.

[28] E. M. Aylward, P. A. Parrilo, and J.-J. E. Slotine. Stability and robustness analysis of nonlinear systems via contraction metrics and sos programming. *Automatica*, 44(8):2163–2170, 2008.

[29] A. Bacciotti and L. Rosier. *Liapunov Functions and Stability in Control Theory*. Springer Science & Business Media, 2006.

[30] J. Baillieul. The geometry of homogeneous polynomial dynamical systems. *Nonlinear Analysis, Theory, Methods and Applications*, 4(5):879–900, 1980.

[31] J. A. Baker. Integration over spheres and the divergence theorem for balls. *The American Mathematical Monthly*, 104(1):36–47, 1997.

[32] D. Bampou and D. Kuhn. Scenario-free stochastic programming with polynomial decision rules. In *Proceedings of the IEEE Conference on Decision and Control*, pages 7806–7812, 2011.

[33] D. Bampou and D. Kuhn. Polynomial approximations for continuous linear programs. *SIAM Journal on Optimization*, 22(2):628–648, 2012.

[34] J. Banks, S. Mohanty, and P. Raghavendra. Local statistics, semidefinite programming, and community detection. *Preprint available at arXiv:1911.01960*, 2019.

[35] R. Bellman. The stability of solutions of linear differential equations. *Duke Mathematical Journal*, 10(4):643–647, 1943.

[36] R. Bellman. Bottleneck problems and dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 39(9):947–951, 1953.

[37] M. A. Ben Sassi and S. Sankaranarayanan. Stability and stabilization of polynomial dynamical systems using Bernstein polynomials. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, pages 291–292. ACM, 2015.

[38] E. Bernuau, D. Efimov, W. Perruquetti, and A. Polyakov. On an extension of homogeneity notion for differential inclusions. In *Proceedings of the European Control Conference*, pages 2204–2209, 2013.

[39] R. Berr and T. Wörmann. Positive polynomials on compact sets. *Manuscripta Mathematica*, 104(2):135–143, 2001.

[40] D. Bertsimas, D. A. Iancu, and P. A. Parrilo. A hierarchy of near-optimal policies for multistage adaptive optimization. *IEEE Transactions on Automatic Control*, 56(12):2809–2824, 2011.

[41] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. In *Springer handbook of robotics*, pages 1371–1394. Springer, 2008.

[42] P. Biswas, T.-C. Lian, T.-C. Wang, and Y. Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networks (TOSN)*, 2(2):188–220, 2006.

[43] F. Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, 1999.

[44] G. Blekherman. Convex forms that are not sums of squares. *arXiv preprint available at arXiv:0910.0656*, 2009.

[45] G. Blekherman. Nonnegative polynomials and sums of squares. *J. Amer. Math. Soc.*, 25(3):617–635, 2012.

[46] G. Blekherman, P. A. Parrilo, and R. Thomas. *Semidefinite Optimization and Convex Algebraic Geometry*. SIAM Series on Optimization, 2013.

[47] V. D. Blondel and A. Megretski, editors. *Unsolved Problems in Mathematical Systems and Control Theory*. Princeton University Press, 2004.

[48] B. Borchers. CSDP, a C library for semidefinite programming. *Optimization Methods and Software*, 11(1-4):613–623, 1999.

[49] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities In System And Control Theory*. SIAM, 1994.

[50] R. N. Buie and J. Abrham. Numerical solutions to continuous linear programming problems. *Zeitschrift für Operations Research*, 17(3):107–117, 1973.

[51] S. O. Chan. Approximation resistance from pairwise-independent subgroups. *Journal of the ACM (JACM)*, 63(3):1–32, 2016.

[52] C.-A. Cheng and H.-P. Huang. Learn the Lagrangian: A vector-valued RKHS approach to identifying Lagrangian systems. *IEEE Transactions on Cybernetics*, 46(12):3247–3258, 2015.

[53] G. Chesi. LMI techniques for optimization over polynomials in control: a survey. *IEEE Transactions on Automatic Control*, 55:2500–2510, 2010.

[54] M.-D. Choi. Positive semidefinite biquadratic forms. *Linear Algebra and its Applications*, 12(2):95–100, 1975.

[55] M.-D. Choi, T.-Y. Lam, and B. Reznick. Real zeros of positive semidefinite forms. I. *Mathematische Zeitschrift*, 171(1):1–26, 1980.

[56] A. Cima and J. Llibre. Algebraic and topological classification of the homogeneous cubic vector fields in the plane. *Journal of Mathematical Analysis and Applications*, 147(2):420–448, 1990.

[57] C. B. Collins. Algebraic classification of homogeneous polynomial vector fields in the plane. *Japan Journal of Industrial and Applied Mathematics*, 13(1):63, 1996.

[58] C. W. Commander. *Optimization problems in telecommunications with military applications.* PhD Thesis, University of Florida, 2007.

[59] C. W. Commander, P. M. Pardalos, V. Ryabchenko, S. Uryasev, and G. Zrazhevsky. The wireless network jamming problem. *Journal of Combinatorial Optimization*, 14(4):481–498, 2007.

[60] C. W. Commander, P. M. Pardalos, V. Ryabchenko, O. Shylo, S. Uryasev, and G. Zrazhevsky. Jamming communication networks under complete uncertainty. *Optimization Letters*, 2(1):53–70, 2008.

[61] P. Comon, G. Golub, L.-H. Lim, and B. Mourrain. Symmetric tensors and symmetric tensor rank. *Preprint available at arXiv:0802.1681*, 2008. URL http://arxiv.org/abs/0802.1681.

[62] E. Coumans and Y. Bai. pybullet, a python module for physics simulation, games, robotics and machine learning. http://pybullet.org/, 2016–2017.

[63] M. Curmei and G. Hall. Shape-constrained regression using sum of squares polynomials. *Preprint available at arXiv:2004.03853*, 2020.

[64] N. da Costa and F. A. Doria. Undecidability and incompleteness in classical mechanics. *International Journal of Theoretical Physics*, 30(8):1041–1073, 1991.

[65] H. Dai, A. Majumdar, and R. Tedrake. Synthesis and optimization of force closure grasps via sequential semidefinite programming. In *Robotics Research*, pages 285–305. Springer, 2018.

[66] E. de Klerk. *Aspects of Semidefinite Programming: Interior Point Algorithms and Selected Applications.* Springer, 2002.

[67] H. Dette and W. J. Studden. Matrix measures, moment spaces and Favard's theorem for the interval [0,1] and [0, ∞). *Linear Algebra and its Applications*, 345(1-3):169–193, 2002.

[68] L. Ding, A. Yurtsever, V. Cevher, J. A. Tropp, and M. Udell. An optimal-storage approach to semidefinite programming using approximate complementarity. *Preprint available at arXiv:1902.03373*, 2019.

[69] F. A. Doria and N. C. A. da Costa. On Arnold's Hilbert symposium problems. In *Computational Logic and Proof Theory*, Lecture Notes in Computer Science, pages 152–158. Springer, 1993.

[70] A. D. Dragan and S. S. Srinivasa. *Formalizing assistive teleoperation*. MIT Press, 2012.

[71] D. Efimov, R. Ushirobira, J. A. Moreno, and W. Perruquetti. On numerical construction of homogeneous Lyapunov functions. In *Proceedings of the IEEE Conference on Decision and Control*, pages 4896–4901, 2017. doi: 10.1109/CDC.2017.8264383.

[72] R. Ehrenborg and G.-C. Rota. Apolarity and canonical forms for homogeneous polynomials. *European Journal of Combinatorics*, 14(3):157–181, 1993.

[73] D. Eisenbud, M. Green, and J. Harris. Cayley-Bacharach theorems and conjectures. *Bulletin of the American Mathematical Society*, 33(3):295–324, 1996.

[74] B. El Khadir, J. Varley, and V. Sindhwani. Teleoperator imitation with continuous-time safety. 2019.

[75] B. Eröcal and W. Stein. The Sage project: Unifying free mathematical software to create a viable alternative to Magma, Maple, Mathematica and MATLAB. In *International Congress on Mathematical Software*, pages 12–27. Springer, 2010.

[76] H. R. Feyzmahdavian, T. Charalambous, and M. Johansson. Asymptotic stability and decay rates of homogeneous positive systems with bounded and unbounded delays. *SIAM Journal on Control and Optimization*, 52(4):2623–2650, 2014.

[77] E. Fischer. Über die Differentiationsprozesse der Algebra. *Journal für die reine und angewandte Mathematik*, 148:1–78, 1918.

[78] L. Fleischer and J. Sethuraman. Efficient algorithms for separated continuous linear programs: the multicommodity flow problem with holding costs and extensions. *Mathematics of Operations Research*, 30(4):916–938, 2005.

[79] T. Fong and C. Thorpe. Vehicle teleoperation interfaces. *Autonomous robots*, 11(1):9–18, 2001.

[80] D. J. Foster, A. Rakhlin, and T. Sarkar. Learning nonlinear dynamical systems from a single trajectory. *Preprint available at arXiv:2004.14681*, 2020.

[81] K. Gatermann and P. A. Parrilo. Symmetry groups, semidefinite programs, and sums of squares. *Journal of Pure and Applied Algebra*, 192(1-3):95–128, 2004.

[82] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[83] K. Goldberg, M. Mascha, S. Gentner, N. Rothenberg, C. Sutter, and J. Wiegley. Desktop teleoperation via the world wide web. In *Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on*, volume 1, pages 654–659. IEEE, 1995.

[84] B. L. Gorissen and D. den Hertog. Approximating the Pareto set of multiobjective linear programs via robust optimization. *Operations Research Letters*, 40(5):319–324, 2012.

[85] S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, pages 3240–3249, 2019.

[86] R. C. Grinold. Continuous programming part one: linear objectives. *Journal of Mathematical Analysis and Applications*, 28(1):32–51, 1969.

[87] L. Grüne. Homogeneous state feedback stabilization of homogeneous systems. In *Proceedings of the 39th IEEE Conference on Decision and Control*, 2000.

[88] W. Hahn. *Stability of Motion*. Springer-Verlag, New York, 1967.

[89] G. Hall. *Optimization over nonnegative and convex polynomials with and without semidefinite programming*. PhD thesis, Princeton University, 2018.

[90] G. Hall. Engineering and business applications of sum of squares polynomials. *Preprint available at arXiv:1906.07961*, 2019.

[91] K. G. Hare, I. D. Morris, N. Sidorov, and J. Theys. An explicit counterexample to the Lagarias–Wang finiteness conjecture. *Advances in Mathematics*, 226(6): 4667–4701, 2011.

[92] D. Hart, E. Shochat, and Z. Agur. The growth law of primary breast cancer as inferred from mammography screening trials data. *British Journal of Cancer*, 78(3):382–387, 1998.

[93] J. W. Helton and J. Nie. Semidefinite representation of convex sets. *Mathematical Programming*, 122(1):21–64, 2010.

[94] D. Henrion and G. Chesi. Guest editorial: special issue on positive polynomials in control. *IEEE Transactions on Automatic Control*, 54(5):935–936, 2009.

[95] D. Henrion and A. Garulli, editors. *Positive polynomials in control*. Lecture Notes in Control and Information Sciences. Springer, 2005.

[96] D. Henrion, S. Naldi, and M. S. El Din. Exact algorithms for linear matrix inequalities. *SIAM Journal on Optimization*, 26(4):2512–2539, 2016.

[97] D. Hilbert. Über die Darstellung Definiter Formen als Summe von Formenquadraten. *Mathematische Annalen*, 32(3):342–350, 1888.

[98] D. Hilbert. Ein Beitrag zur Theorie des Legendre'schen Polynoms. *Acta Mathematica*, 18:155–159, 1894.

[99] P. Hilton and J. Pedersen. Catalan numbers, their generalization, and their uses. *The Mathematical Intelligencer*, 13(2):64–75, Mar 1991. ISSN 0343-6993.

[100] L. Huber, A. Billard, and J.-J. E. Slotine. Avoidance of convex and concave obstacles with convergence ensured through contraction. *IEEE Robotics and Automation Letters*, PP:1–1, 2019.

[101] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 25(2):328–373, 2013.

[102] Z. Jarvis-Wloszek, R. Feeley, W. Tan, K. Sun, and A. Packard. Some controls applications of sum of squares programming. In *Proceedings of the IEEE Conference on Decision and Control*, volume 5, pages 4676–4681, 2003.

[103] M. A. H. H. Jerbia. The stabilization of homogeneous cubic vector fields in the plane. *Applied Mathematics Letters*, 7(4):95–99, 1994.

[104] J. Jouffroy and T. I. Fossen. A tutorial on incremental stability analysis using contraction theory. *Modeling, Identification and control*, 31(3):93, 2010.

[105] R. Jungers. *The Joint Spectral Radius: Theory and Applications*. Lecture Notes in Control and Information Sciences. Springer-Verlag, 2009.

[106] R. M. Jungers and V. D. Blondel. On the finiteness property for rational matrices. *Linear Algebra and its Applications*, 428(10):2283–2295, 2008.

[107] R. Kamyar and M. Peet. Polynomial optimization with applications to stability analysis and control—alternatives to sum of squares. *Discrete and Continuous Dynamical Systems*, 2014.

[108] R. Kamyar, M. M. Peet, and Y. Peet. Solving large-scale robust stability problems by exploiting the parallel structure of Polya's theorem. *IEEE Transactions on Automatic Control*, 58(8):1931–1947, 2013.

[109] D. Kappler, F. Meier, J. Issac, J. Mainprice, C. G. Cifuentes, M. Wüthrich, V. Berenz, S. Schaal, N. Ratliff, and J. Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018.

[110] M. Kawski. Stability and nilpotent approximations. In *Proceedings of the 27th IEEE Conference on Decision and Control*, pages 1244–1248, 1988.

[111] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

[112] H. Khalil. *Nonlinear Systems*. Prentice Hall, 2002. Third edition.

[113] S. M. Khansari-Zadeh and A. Billard. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27(5): 943–957, 2011.

[114] S. M. Khansari-Zadeh and A. Billard. A dynamical system approach to realtime obstacle avoidance. *Autonomous Robots*, 32(4):433–454, 2012.

[115] S. M. Khansari-Zadeh and A. Billard. Learning control Lyapunov function to ensure stability of dynamical system-based robot reaching motions. *Robotics and Autonomous Systems*, 6(62), 2014.

[116] S. M. Khansari-Zadeh and O. Khatib. Learning potential functions from human demonstrations with encapsulated dynamic and compliant behaviors. *Autonomous Robots*, 41(1):45–69, 2017.

[117] O. Khatib. Real-time obstacle avoidance for manipulators and mobile robots. *The international journal of robotics research*, 5(1):90–98, 1986.

[118] M. Kojima. Sums of squares relaxations of polynomial semidefinite programs. *Research report B-397, Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology*, 2003.

[119] J. Z. Kolter and G. Manek. Learning stable deep dynamics models. In *Advances in Neural Information Processing Systems*, pages 11126–11134, 2019.

[120] J. C. Lagarias and Y. Wang. The finiteness conjecture for the generalized spectral radius of a set of matrices. *Linear Algebra and its Applications*, 214: 17–42, 1995.

[121] J. Lasserre and J. Hiriart-Urruty. Mathematical properties of optimization problems defined by positively homogeneous functions. *Journal of Optimization Theory and Applications*, 112(1):31–52, 2002.

[122] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[123] J. B. Lasserre. A "joint+marginal" approach to parametric polynomial optimization. *SIAM Journal on Optimization*, 20, 2009.

[124] J. B. Lasserre. *Moments, Positive Polynomials And Their Applications*, volume 1. World Scientific, 2010.

[125] J. B. Lasserre. *Moments, Positive Polynomials and Their Applications*. World Scientific, 2010.

[126] M. Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

[127] R. S. Lehman. On the continuous simplex method. Technical Report RM-1386, Rand Corporations, Santa Monica., 1954.

[128] A. Lemme, Y. Meirovitch, S. M. Khansari-Zadeh, T. Flash, A. Billard, and J. J. Steil. Open-source benchmarking for learned reaching motion generation in robotics. 2015.

[129] T. Leth, R. Wisniewski, and C. Sloth. On the existence of polynomial Lyapunov functions for rationally stable vector fields. In *Proceedings of the IEEE Conference on Decision and Control*, pages 4884–4889, 2017.

[130] N. Levinson. A class of continuous linear programming problems. *Journal of Mathematical Analysis and Applications*, 16(1):73–83, 1966.

[131] J. Löfberg. Yalmip : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, 2004.

[132] J. Löfberg and P. A. Parrilo. From coefficients to samples: a new approach to SOS optimization. In *Proceedings of the IEEE Conference on Decision and Control*, volume 3, pages 3154–3159, 2004.

[133] W. Lohmiller and J.-J. E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

[134] F. Lukács. Verschärfung des ersten Mittelwertsatzes der Integralrechnung für rationale Polynome. *Mathematische Zeitschrift*, 2(3):295–305, 1918.

[135] X. Luo and D. Bertsimas. A new algorithm for state-constrained separated continuous linear programs. *SIAM Journal on Control and Optimization*, 37 (1):177–210, 1998.

[136] V. Magron, D. Henrion, and J. B. Lasserre. Approximating Pareto curves using semidefinite relaxations. *Operations Research Letters*, 42(6):432–437, 2014.

[137] A. Majumdar, A. A. Ahmadi, and R. Tedrake. Control design along trajectories with sums of squares programming. In *International Conference on Robotics and Automation*, pages 4054–4061. IEEE, 2013.

[138] A. Majumdar, G. Hall, and A. A. Ahmadi. A survey of recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics. *Preprint available at arXiv:1908.05209*, 2019.

[139] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[140] N. Marturi, A. Rastegarpanah, C. Takahashi, M. Adjigble, R. Stolkin, S. Zurek, M. Kopicki, M. Talha, J. A. Kuo, and Y. Bekiroglu. Towards advanced robotic manipulation for nuclear decommissioning: a pilot study on tele-operation and autonomy. In *International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*, pages 1–8. IEEE, 2016.

[141] P. Min. Binvox, a 3d mesh voxelizer, 2004.

[142] L. Moreau, D. Aeyels, J. Peuteman, and R. Sepulchre. Homogeneous systems: stability, boundedness and duality. In *Proceedings of the 14th Symposium on Mathematical Theory of Networks and Systems*, 2000.

[143] T. Motzkin. The arithmetic-geometric inequality. *In Proceedings of Symposium on Inequalities*, pages 205–224, 1967.

[144] K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming, 39:117–129*, 1985.

[145] M. Nagumo. Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen. *Proceedings of the Physico-Mathematical Society of Japan.*, 24:551–559, 1942.

[146] Y. Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000.

[147] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. SCS: Splitting conic solver, version 2.1.0. https://github.com/cvxgrp/scs, Nov. 2017.

[148] Z. J.-W. R. F. W. T. K. S. A. Packard. Some controls applications of sum of squares programming. In *Proceedings of the IEEE Conference on Decision and Control*, volume 5, pages 4676–4681, 2003.

[149] A. Papachristodoulou and S. Prajna. On the construction of Lyapunov functions using the sum of squares decomposition. In *Proceedings of the IEEE Conference on Decision and Control*, volume 3, pages 3482–3487, 2002.

[150] A. Papachristodoulou, P. A. Parrilo, P. Seiler, J. Anderson, G. Valmorbida, S. Prajna, and P. Seiler. *SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB*. 2013.

[151] M. M. P. A. Papachristodoulou. A converse sum of squares Lyapunov result with a degree bound. *IEEE Transactions on Automatic Control*, 57(9):2281–2293, 2012.

[152] D. Papp. Semi-infinite programming using high-degree polynomial interpolants and semidefinite programming. *SIAM Journal on Optimization*, 27(3):1858–1879, 2017.

[153] D. Papp and F. Alizadeh. Semidefinite characterization of sum-of-squares cones in algebras. *SIAM Journal on Optimization*, 23(3):1398–1423, 2013.

[154] D. Papp and S. Yildiz. Sum-of-squares optimization without semidefinite programming. *SIAM Journal on Optimization*, 29(1):822–851, 2019.

[155] A. A. A. P. A. Parrilo. Converse results on existence of sum of squares Lyapunov functions. In *Proceedings of the IEEE Conference on Decision and Control*, pages 6516–6521, 2011.

[156] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, May 2000.

[157] P. A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003.

[158] P. A. Parrilo and B. Sturmfels. Minimizing polynomial functions. *Algorithmic and Quantitative Real Algebraic Geometry, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 60:83–99, 2003.

[159] E. Pauwels, D. Henrion, and J.-B. B. Lasserre. Inverse optimal control with polynomial optimization. *Preprint available at arXiv:1403.5180*, 2014.

[160] M. M. Peet. Exponentially Stable Nonlinear Systems Have Polynomial Lyapunov Functions on Bounded Regions. *IEEE Transactions on Automatic Control*, 54(5):979–987, May 2009. ISSN 0018-9286. doi: 10.1109/TAC.2009.2017116.

[161] M. M. Peet and A. Papachristodoulou. A converse sum of squares Lyapunov result with a degree bound. *IEEE Transactions on Automatic Control*, 57(9): 2281–2293, 2012.

[162] A. F. Perold. Fundamentals of a continuous time simplex method. Technical Report SOL-78-26, Rand Corporations, Santa Monica., 1978.

[163] L. Porkolab and L. Khachiyan. On the complexity of semidefinite programs. *Journal of Global Optimization*, 10(4):351–365, 1997.

[164] M. Posa, T. Koolen, and R. Tedrake. Balancing and step recovery capturability via sums-of-squares optimization. In *Robotics: Science and Systems*, pages 12–16, 2017.

[165] A. P. S. Prajna. On the construction of Lyapunov functions using the sum of squares decomposition. In *Proceedings of the IEEE Conference on Decision and Control*, 2002.

[166] S. Prajna, A. Papachristodoulou, and P. A. Parrilo. *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, 2002. Available from http://www.cds.caltech.edu/sostools and http://www.mit.edu/~parrilo/sostools.

[167] J. B. Prolla and C. S. Guerreiro. An extension of Nachbin's theorem to differentiable functions on Banach spaces with the approximation property. *Arkiv för Matematik*, 14(1-2):251, 1976.

[168] M. C. Pullan. An algorithm for a class of continuous linear programs. *SIAM Journal on Control and Optimization*, 31(6):1558–1577, 1993.

[169] M. C. Pullan. Forms of optimal solutions for separated continuous linear programs. *SIAM Journal on Control and Optimization*, 33(6):1952–1977, 1995.

[170] M. C. Pullan. A duality theory for separated continuous linear programs. *SIAM Journal on Control and Optimization*, 34(3):931–965, 1996.

[171] M. C. Pullan. Convergence of a general class of algorithms for separated continuous linear programs. *SIAM Journal on Optimization*, 10(3):722–731, 2000.

[172] M. Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42(3):969–984, 1993.

[173] H. Ravichandar, I. Salehi, and A. Dani. Learning partially contracting dynamical systems from demonstrations. In *Conference on Robot Learning (CoRL)*, 2017.

[174] B. Reznick. Some concrete aspects of Hilbert's 17th problem. *Contemporary Mathematics*, 253:251–272, 2000.

[175] B. Reznick. On Hilbert's construction of positive polynomials. *Preprint available at arXiv:0707.2156*, 2007.

[176] B. Reznick. Blenders. In *Notions of Positivity and the Geometry of Polynomials*, pages 345–373. Springer, 2011.

[177] R. M. Robinson. Some definite polynomials which are not sums of squares of real polynomials. In *Notices of the American Mathematical Society*, volume 16, page 554, 1969.

[178] A. B. L. Rosier. *Liapunov Functions and Stability in Control Theory*. Springer, 2005.

[179] L. Rosier. Homogeneous Lyapunov function for homogeneous continuous vector fields. *Systems and Control Letters*, 19(6):467–473, 1992. ISSN 0167-6911.

[180] R. Sachs, L. Hlatky, and P. Hahnfeldt. Simple ODE models of tumor growth and anti-angiogenic or radiation treatment. *Mathematical and Computer Modelling*, 33(12-13):1297–1305, 2001.

[181] N. Samardzija. Stability properties of autonomous homogeneous polynomial differential systems. *Journal of Differential Equations*, 48(1):60–70, 1983.

[182] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

[183] H. Schaeffer, G. Tran, R. Ward, and L. Zhang. Extracting structured dynamical systems using sparse optimization with very few samples. *Preprint available at arXiv:1805.04158*, 2018.

[184] C. Scheiderer. A Positivstellensatz for projective real varieties. *Manuscripta Mathematica*, 138(1-2):73–88, 2012.

[185] C. Scheiderer. An observation on positive definite forms. Preprint available at arXiv:1602.03986, 2016.

[186] C. W. Scherer and C. W. Hol. Matrix sum-of-squares relaxations for robust semi-definite programs. *Mathematical Programming*, 107(1-2):189–211, 2006.

[187] A. Seidenberg. A new decision method for elementary algebra. *Annals of Mathematics*, 60(2):365–374, 1954.

[188] A. Shapiro. On duality theory of conic linear problems. In *Semi-Infinite Programming: Recent Advances*, pages 135–165. Springer US, Boston, MA, 2001.

[189] V. Sindhwani, S. Tu, and M. Khansari. Learning contracting vector fields for stable imitation learning. *Preprint available at arXiv:1804.04878*, 2018.

[190] S. Singh, V. Sindhwani, J.-J. Slotine, and M. Pavone. Learning stabilizable dynamical systems via control contraction metrics. In *Workshop on Algorithmic Foundations of Robotics*, 2018.

[191] S. Singh, S. M. Richards, V. Sindhwani, J.-J. E. Slotine, and M. Pavone. Learning stabilizable nonlinear dynamics with contraction-based regularization. *Preprint available at arXiv:1907.13122*, 2019.

[192] P. Skehan. On the normality of growth dynamics of neoplasms in vivo: a data base analysis. *Growth*, 50(4):496—515, 1986. ISSN 0017-4793. URL http://europepmc.org/abstract/MED/3596327.

[193] H. L. Smith. *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*. Number 41. American Mathematical Soc., 2008.

[194] M. H. Stone. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21(5):237–254, 1948.

[195] M. Talamini, K. Campbell, and C. Stanfield. Robotic gastrointestinal surgery: early experience and system description. *Journal of laparoendoscopic & advanced surgical techniques*, 12(4):225–232, 2002.

[196] A. Tarski. A decision method for elementary algebra and geometry. In *Quantifier Elimination and Cylindrical Algebraic Decomposition*, Texts and Monographs in Symbolic Computation, pages 24–84. Springer, 1998.

[197] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario. Medical robotics and computer-integrated surgery. In *Springer handbook of robotics*, pages 1657–1684. Springer, 2016.

[198] A. F. Timan. *Theory of Approximation of Functions of a Real Variable*. Elsevier, 2014.

[199] K. C. Toh, R. H. Tütüncü, and M. J. Todd. *SDPT3 - a MATLAB software package for semidefinite-quadratic-linear programming*. URL http://www.math.cmu.edu/~reha/sdpt3.html.

[200] W. F. Tyndall. A duality theorem for a class of continuous linear programming problems. *Journal of the Society for Industrial and Applied Mathematics*, 13 (3):644–666, 1965.

[201] W. F. Tyndall. An extended duality theorem for continuous linear programming problems. *SIAM Journal on Applied Mathematics*, 15(5):1294–1298, 1967.

[202] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38 (1):49–95, 1996.

[203] A. V. E. Veretennikova. On partial derivatives of multivariate Bernstein polynomials. *Siberian Advances in Mathematics*, 26(4):294–305, 2016.

[204] C. Walkden. Lecture notes on Ergodic theory. *The University of Manchester*, 2018.

[205] X. Wang, S. Zhang, and D. Yao. Separated continuous conic programming: strong duality and an approximation algorithm. *SIAM Journal on Control and Optimization*, 48(4):2118–2138, 2009.

[206] R. Washington, K. Golden, J. Bresina, D. E. Smith, C. Anderson, and T. Smith. Autonomous rovers for mars exploration. In *Aerospace Conference, 1999. Proceedings. 1999 IEEE*, volume 1, pages 237–251. IEEE, 1999.

[207] G. Weiss. A simplex based algorithm to solve separated continuous linear programs. *Mathematical Programming*, 115(1):151–198, 2008.

[208] T. Weisser, B. Legat, C. Coey, L. Kapelevich, and J. P. Vielma. Polynomial and moment optimization in Julia and JuMP. In *JuliaCon*, 2019.

[209] Y. Yang, K. Caluwaerts, A. Iscen, T. Zhang, J. Tan, and V. Sindhwani. Data efficient reinforcement learning for legged robots. In *Conference on Robot Learning*, pages 1–10, 2020.

[210] T. Zhang, Z. McCarthy, O. Jowl, D. Lee, X. Chen, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[211] Y. Zheng, G. Fantuzzi, and A. Papachristodoulou. Exploiting sparsity in the coefficient matching conditions in sum-of-squares programming using ADMM. *IEEE Control Systems Letters*, 1(1):80–85, 2017.

[212] V. I. Zubov. Methods of A.M. Lyapunov and their applications. *The Netherlands*, 1964.