Learning Dynamical Systems with Side Information (short version)

Amir Ali Ahmadi AAA@PRINCETON.EDU*

Bachir El Khadir BKHADIR@PRINCETON.EDU*

Abstract

We present a mathematical formalism and a computational framework for the problem of learning a dynamical system from noisy observations of a few trajectories and subject to *side information* (e.g., physical laws or contextual knowledge). We identify six classes of side information which can be imposed by semidefinite programming and that arise naturally in many applications. We demonstrate their value on two examples from epidemiology and physics. Some density results on polynomial dynamical systems that either exactly or approximately satisfy side information are also presented.

Keywords: Learning, Dynamical Systems, Sum of Squares Optimization, Semidefinite Programming

1. Introduction

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.

In this paper, we present a mathematical formalism of the problem of learning a dynamical system with side information. We identify a list of six notions of side information that are commonly encountered in practice and can be enforced in any combination by semidefinite programming (SDP). After presenting these notions in Section 2.1, we describe the SDP formulation in Section 3, demonstrate the applicability of the approach on two examples in Section 4, and end with theoretical justification of our methodology in Section 5.

2. Problem Formulation

Our interest in this paper is to learn a dynamical system

$$\dot{x}(t) = f(x(t)), \quad f : \Omega \to \mathbb{R}^n, \tag{1}$$

over a given compact set $\Omega \subset \mathbb{R}^n$ from noisy observations of a limited number of its trajectories. We assume that the unknown vector field f is continuously differentiable ($f \in C^1$ for short). This assumption is often met in applications, and is known to be a sufficient condition for existence and

* This work was partially supported by the MURI award of the AFOSR, the DARPA Young Faculty Award, the CAREER Award of the NSF, the Google Faculty Award, the Innovation Award of the School of Engineering and Applied Sciences at Princeton University, and the Sloan Fellowship.

uniqueness of solutions to (1) (see, e.g., [10]). In our setting, we have access to a set of the form

$$\mathcal{D} := \{(x_i, y_i), \quad i = 1, \ldots, N\}, \tag{2}$$

where $x_i \in \Omega$ is a possibly noisy measurement of the state of the dynamical system, and $y_i \in \mathbb{R}^n$ is a noisy measurement of $f(x_i)$. Typically, this training set is obtained from observation of a few trajectories of (1). The vectors y_i could be either directly accessible (e.g., from sensor measurements) or approximated using a finite-difference scheme on the state variables.

Finding a vector field $f_{\mathcal{F}}$ that best agrees with the unknown vector field f among a particular subspace $\mathcal{F}$ of continuously-differentiable functions amounts to solving a least-squares problem:

$$f_{\mathcal{F}} \in \arg\min_{p \in \mathcal{F}} \sum_{(x_i, y_i) \in \mathcal{D}} \|p(x_i) - y_i\|^2. \tag{3}$$

While we work with the least-squares loss of simplicity, it turns out that our SDP-based approach can readily handle other types of losses such as the ℓ_1 loss, the ℓ_∞ loss, and any loss given by an *sos-convex* function (see [9] for a definition and also [12, Theorem 3.3]).

2.1. Side information

In addition to consistency with f, we desire for our learned vector field $f_{\mathcal{F}}$ to also generalize well in conditions that were not observed in the training data. Indeed, the optimization problem in (3) only dictates how the candidate vector field should behave on the training data, which could easily lead to over-fitting, especially if the function class $\mathcal{F}$ is large and the observations are limited. Let us demonstrate this issue with a simple example.

Example 1 *Consider the two-dimensional vector field $f(x_1, x_2) := (-x_2, x_1)^T$. The trajectories of the system $\dot{x} = f(x)$ from any initial condition are given by circular orbits. In particular, if started from the point $x_0 := (1, 0)^T$, the trajectory is given by $x(t, x_0) = (\cos(t), \sin(t))^T$. Hence, for any function $g : \mathbb{R}^2 \to \mathbb{R}^2$, the vector field $h(x) := f(x) + (x_1^2 + x_2^2 - 1)g(x)$ agrees with f on the sample trajectory $x(t, x_0)$. However, the behavior of the trajectories of h depend on the arbitrary choice of the function g. If $g(x) = x$ for instance, the trajectories of h starting outside of the unit disk diverge to infinity.*

To address the issues of over-fitting and scarcity of data, 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 vector fields. Our goal is then to replace the optimization problem in (3) with

$$\min_{p \in \mathcal{F} \cap S_1 \cap \cdots \cap S_k} \sum_{(x_i, y_i) \in \mathcal{D}} \|p(x_i) - y_i\|^2, \tag{4}$$

i.e., to find a vector field $p \in \mathcal{F}$ that satisfies the finite list of side information $S_1, \ldots, S_k$ that f is known to satisfy.

For arbitrary side information S_i, it might be unclear how one could solve (4). Below, we identify six types of side information that we believe are useful in practice (see, e.g., Section 4) and can be tackled using semidefinite programming (see Sections 3 and 5).

- **Interpolation at a finite set of points.** For a set of points $\{(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}^n\}_{i=1}^m$, we denote by $\mathbf{Interp}(\{x_i, y_i\}_{i=1}^m)$ the set of vector fields $f \in C^1$ that satisfy $f(x_i) = y_i$ for $i = 1, \ldots, m$. An important special case of this 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.
- **Sign symmetry.** For any two $n \times n$ diagonal matrices A and B with 1 or -1 on the diagonal, we define $\mathbf{Sym}(A, B)$ to be the set of vector fields $f \in C^1$ satisfying the symmetry condition

$f(Ax) = Bf(x)\ \forall x \in \mathbb{R}^n$. If I denotes the $n \times n$ identity matrix, then the set $\mathbf{Sym}(-I, I)$ (resp. $\mathbf{Sym}(-I, -I)$) is exactly the set of even (resp. odd) vector fields.

- **Coordinate nonnegativity.** For any sets $B_i \subseteq \Omega$, $i = 1, \ldots, n$, we denote by $\mathbf{Pos}(\{\trianglerighteq_i, B_i\}_{i=1}^n)$ the set of vector fields $f \in C^1$ that satisfy $f_i(x) \trianglerighteq_i 0\ \forall x \in B_i\ \forall i \in \{1, \ldots, n\}$, where $\trianglerighteq_i$ stands for $\geq$ or $\leq$. These constraints are useful when we know that certain components of the state variables are increasing or decreasing functions of time in some regions of the space.
- **Coordinate directional monotonicity.** For any sets $B_{i,j} \subseteq \Omega$, $i, j = 1, \ldots, n$, we denote the set of vector fields $f \in C^1$ that satisfy $\frac{\partial f_i}{x_j}(x) \trianglerighteq_{i,j} 0\ \forall x \in B_{i,j}\ \forall i, j \in \{1, \ldots, n\}$, where $\trianglerighteq_{ij}$ stands as before for $\geq$ or $\leq$, by $\mathbf{Mon}(\{\trianglerighteq_{i,j}, B_{i,j}\}_{i,j=1}^n)$. An important special case of this is when $B_{ij} = \Omega$ and $\trianglerighteq_{ij}$ is taken to be $\geq$ for all $i \neq j$. In this case, the side information is the knowledge of the following property of the vector field f:
$$\forall x_0, \tilde{x}_0 \in \Omega \quad [x_0 \leq \tilde{x}_0 \implies x(t, x_0) \leq x(t, \tilde{x}_0)\ \forall t \geq 0].$$
Here the inequalities are interpreted elementwise, and the notation $x(t, x_0)$ for example denotes the trajectory of the vector field f starting from the point x_0.
- **Invariance of a set.** We say that a set $B \subseteq \Omega$ is invariant under a vector field $f \in C^1$ if any trajectory of the dynamical system $\dot{x} = f(x)$ which starts in B stays in B forever. In particular, if $B = \{x \in \mathbb{R}^n \mid h_i(x) \geq 0,\ i = 1, , \ldots, m\}$ for some C^1 functions h_i, then invariance of the set B under the vector field f is equivalent to the following constraint:
$$\forall i \in \{1, \ldots, m\}\ \forall x \in B \quad [h_i(x) = 0 \implies \langle f(x), \nabla h_i(x) \rangle \geq 0]. \tag{5}$$
The set of all C^1 vector fields under which the set B is invariant is denoted by $\mathbf{Inv}(B)$.
- **Gradient and Hamiltonian systems.** The vector field $f \in C^1$ is said to be a *gradient* vector field if there exists a scalar-valued function $V : \Omega \to \mathbb{R}$ such that $f(x) = -\nabla V(x)\ \forall x \in \Omega$. Typically, the function V is interpreted as a potential or energy that decreases along the trajectories of the dynamical system $\dot{x} = f(x)$. The set of gradient vector fields is denoted by $\mathbf{Grad}$. A dynamical system is said to be *Hamiltonian* if the dimension n of the state space x is even, and there exists a scalar-valued function $H : \Omega \longrightarrow \mathbb{R}$ such that
$$f_i(p, q) = -\frac{\partial H}{\partial q_i}(p, q) \text{ and } f_{\frac{n}{2}+i}(p, q) = \frac{\partial H}{\partial p_i}(p, q),$$
where $p = (x_1, \ldots, x_{\frac{n}{2}})^T$ and $q = (x_{\frac{n}{2}+1}, \ldots, x_n)^T$. The coordinates p and q are usually called *momentum* and *position* respectively, following terminology from physics. Note that a Hamiltonian system conserves the quantity H along its trajectories. The set of Hamiltonian vector fields is denoted by $\mathbf{Ham}$. For related work on learning Hamiltonian systems, see [7; 2].

3. Learning Polynomial Vector Fields Subject to Side Information

To completely define the optimization problem in (4), we still have to specify the function class $\mathcal{F}$. Among the possible choices are reproducing kernel Hilbert spaces [18; 19; 5], trigonometric functions, and functions parameterized by neural networks [4; 7]. In this paper, we take $\mathcal{F}$ to be the set
$$\mathcal{P}_d := \{p : \mathbb{R}^n \to \mathbb{R}^n \mid p_i \text{ is a (multivariate) polynomial of degree } d \text{ for } i = 1, \ldots, n\}.$$
Furthermore, we assume that the set Ω and all its subsets considered in Section 2.1 in the definitions of side information (i.e., the sets B_i in the definition of $\mathbf{Pos}(\{\trianglerighteq_i, B_i\}_{i=1}^n)$, the sets B_{ij} in the definition of $\mathbf{Mon}(\{\trianglerighteq_{i,j}, B_{i,j}\}_{i,j=1}^n)$, and the set B in the definition of $\mathbf{Inv}(B)$) are *closed basic semi-algebraic*. We recall that a closed basic semialgebraic set is a subset of the Euclidean space of

the form

$$\Lambda := \{x \in \mathbb{R}^n |\ g_i(x) \geq 0, i = 1, \ldots, m\}, \tag{6}$$

where $g_1, \ldots, g_m$ are polynomial functions.

These choices are motivated by two reasons. The first is that polynomial functions are expressive enough to approximate a large family of functions. The second reason which shall be made clear in this paper 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. We note that the problem of fitting a polynomial vector field to data has appeared e.g. in [17], 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 [8, Chapter 8], where similar algebraic techniques are used to impose constraints such as convexity and monotonicity on a polynomial regressor. See also [6] 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.

With our choices, the optimization problem in (4) has as decision variables the coefficients of a candidate polynomial vector field p. The objective function is a convex quadratic function of these coefficients, and the constraints are twofold: (i) affine constraints in the coefficients of p, and (ii) constraints of the form

$$q(x) \geq 0\ \forall x \in \Lambda, \tag{7}$$

where Λ is a given closed basic semialgebraic set of the form (6), and q is a (scalar-valued) polynomial whose coefficients depend affinely on the coefficients of the polynomial p. For example, it is easy to see that membership to $\mathbf{Interp}(\{x_i, y_i\}_{i=1}^m)$, $\mathbf{Sym}(A, B)$, $\mathbf{Grad}$, or $\mathbf{Ham}$ is given by affine constraints, while membership to $\mathbf{Pos}(\{\unrhd_i, B_i\}_{i=1}^n)$, $\mathbf{Mon}(\{\unrhd_{i,j}, B_{i,j}\}_{i,j=1}^n)$, or $\mathbf{Inv}(B)$ can be cast as constraints of the type in (7). 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.

An idea pioneered to a large extent by Lasserre [11] and Parrilo [15] has been to write algebraic sufficient conditions for (7) 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

$$q(x) = \sigma_0(x) + \sum_{i=1}^{m} \sigma_i(x) g_i(x) \tag{8}$$

holds, then, clearly, the constraint in (7) must be satisfied. When the degree of the sos polynomials σ_i is bounded by an integer r, we refer to the identity in (8) as the *degree-r sos certificate* corresponding to the constraint in (7). Conversely, a celebrated result in algebraic geometry [16] 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.

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 (8) can be automated via semidefinite programming. 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., [14]): 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

$h(x) = z(x)^T Q z(x)$, where $z(x)$ here denotes the vector of all monomials in x of degree less than or equal to d.

4. Illustrative Experiments

4.1. Diffusion of a contagious disease

The following dynamical system has appeared in the epidemiology literature (see, e.g., [3]) as a model for the spread of Gonorrhea in a heterosexual population:

$$\dot{x} = f(x), \text{where } x \in \mathbb{R}^2 \text{ and } f(x) = \begin{pmatrix} -a_1 x_1 + b_1(1 - x_1)x_2 \\ -a_2 x_2 + b_2(1 - x_1)x_2 \end{pmatrix}. \tag{9}$$

Here, the quantity $x_1(t)$ (resp. $x_2(t)$) represents the fraction of infected males (resp. females) in the population. The parameters a_i and b_i respectively denote the recovery and infection rates for males when $i = 1$, and for females when $i = 2$. We take $(a_1, b_1, a_2, b_2) = (0.05, 0.1, 0.05, 0.1)$, and we plot the resulting vector field f in Figure 1a. We suppose that this vector field is *unknown* to us, and our goal is to learn it from a few noisy snapshots of a single trajectory. More specifically, we have access to the training data set

$$\mathcal{D} := \left\{ \left(x(t_i, x_0), f(x(t_i, x_0)) + 10^{-4} \begin{pmatrix} \varepsilon_i^1 \\ \varepsilon_i^2 \end{pmatrix} \right) \right\}_{i=1}^{20},$$

where $x(t, x_0)$ is the trajectory obtained when the flow in (9) is started from the initial condition $x_0 = (0.7, 0.3)^T$, the scalars $t_i := i/20$ represent a uniform subdivision of the time interval $[0, 1]$, and the scalars $\varepsilon_i^1, \varepsilon_i^2$ are independent standard normal variables.

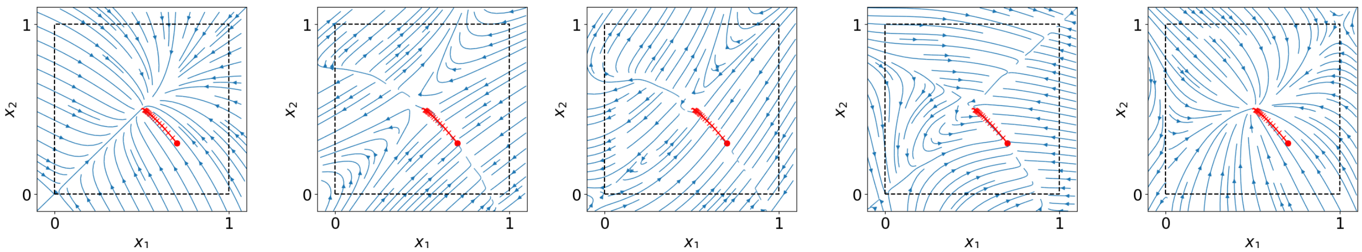
Following our approach in Section 3, we parameterize our candidate vector field $p : \mathbb{R}^2 \to \mathbb{R}^2$ as a polynomial of degree d. Note that the true vector field f is a polynomial of degree 2. In this experiment, we pretend that f is unknown to us and consider an over-parameterized model of the true dynamics by taking $d = 3$. In absence of any side information, one could solve the least-squares problem

$$\min_{p \in \mathcal{P}_3} \sum_{(x_i, y_i) \in \mathcal{D}} \|p(x_i) - y_i\|^2 \tag{10}$$

to find a cubic polynomial that best agrees with the training data. The solution to problem (10) is plotted in Figure 1b. Observe that while the learned vector field replicates the behavior of the vector field f on the observed trajectory, it differs significantly from f on the rest of the unit box. To remedy this problem, we leverage the following side information that are available from the context without knowing the exact structure of f.

- **Equilibrium point at the origin (Interp).** The disease cannot spread if no male or female is infected. This side information corresponds to our vector field p having an equilibrium point at the origin, i.e., $p(0, 0) = 0$. We simply add this linear constraint to problem (10) and plot the resulting vector field in Figure 1c. Note from Figure 1b that the least-squares solution does not satisfy this side information.

- **Invariance of the box $[0, 1]^2$ (Inv).** The state variables (x_1, x_2) of the dynamics in (9) represent fractions, and as such, the vector $x(t)$ should be contained in the box $[0, 1]^2$ at all times $t \geq 0$. Mathematically, this corresponds to the four (univariate) polynomial nonnegativity constraints
$$p_2(x_1, 0) \geq 0, p_2(x_1, 1) \leq 0 \ \forall x_1 \in [0, 1], \quad p_1(0, x_2) \geq 0, p_1(1, x_2) \leq 0 \ \forall x_2 \in [0, 1],$$
which 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 (8). For instance, we replace the constraint $p_2(x_1, 0) \geq 0 \ \forall x_1 \in [0, 1]$ with the linear constraints obtained from equating the coefficients of the two sides of the polynomial identity $p_2(x_1, 0) = x_1 s_0(x_1) +$



(a) The vector field in (9) (b) no side information (c) **Interp** (d) **Interp** $\cap$ **Inv** (e) **Interp** $\cap$ **Inv** $\cap$ **Mon**

Figure 1: (Figure 1a) Streamplot of the true and unknown vector field in (9) that is to be learned from a single trajectory starting from $(0.7, 0.3)^T$. (Figures 1b to 1e) Streamplots of the polynomial vector fields of degree 3 returned by our SDPs as more side information constraints are added. In each case, the trajectory of the learned vector field starting from $(0.7, 0.3)^T$ is also plotted.

$(1 - x_1)s_1(x_1)$. Here, the new decision variables s_0 and s_1 are (univariate) quadratic polynomials that are constrained to be sos. Obviously, this algebraic identity is sufficient for nonnegativity of $p_2(x_1, 0)$ over $[0, 1]$; In this case, it also happens to be necessary [13]. The output of the SDP which imposes the invariance of the unit box and the equilibrium at the origin is plotted in Figure 1d.

- **Coordinate directional monotonicity (Mon).** We expect that if the fraction of males infected rises in the population, the rate of infection of females should increase. Mathematically, this amounts to the constraint that $\frac{\partial p_2}{\partial x_1}(x) \geq 0 \ \forall x \in [0, 1]^2$. Similarly, by changing the roles played by males and females, we obtain the constraint $\frac{\partial p_1}{\partial x_2}(x) \geq 0 \ \forall x \in [0, 1]^2$. Note that $[0, 1]^2$ is a closed basic semialgebraic set, so in the same spirit as the previous bullet point, we replace each one of these constraints with its corresponding degree-2 sos certificate (see (8)). The resulting vector field is plotted in Figure 1e.

Note from Figures 1b to 1e 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 Figure 1e is quite similar qualitatively to the truth in Figure 1a even though only a single noisy trajectory is used for learning.

4.2. The simple pendulum

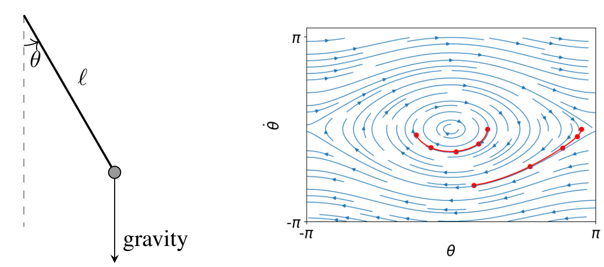


Figure 2: The simple pendulum and its phase portrait.

In this subsection, we consider the simple pendulum system, i.e., a mass m hanging from a massless rod of length ℓ (see Figure 2). 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 $\ddot{\theta} = -g/\ell \sin \theta$ for the pendulum may be obtained, where g is is the local acceleration of gravity. This is a one-dimensional second-order system that we convert to a first-order system as follows:

$$\dot{x} = \begin{pmatrix} \dot{\theta} \\ \ddot{\theta} \end{pmatrix} = f(\theta, \dot{\theta}) := \begin{pmatrix} \dot{\theta} \\ -\frac{g}{\ell} \sin \theta \end{pmatrix}. \tag{11}$$

We take the vector field in (11) to be the ground truth with $g = \ell = 1$, and we observe from it a noisy version of two trajectories $x(t, x_0)$ and $x(t, \tilde{x}_0)$ sampled at times $t_i = 1/5, 2/5, \ldots, 1$, with $x_0 = (\frac{\pi}{4}, 0)^T$ and $\tilde{x}_0 = (\frac{9\pi}{10}, 0)^T$ (see Figure 2). More precisely, we assume that we have the following training data set:

$$\mathcal{D} := \left\{(\theta(t_i, x_0), \dot{\theta}(t_i, x_0), \ddot{\theta}(t_i, x_0)) + 10^{-2}\varepsilon_i^1\right\}_{i=1}^5 \bigcup \left\{(\theta(t_i, \tilde{x}_0), \dot{\theta}(t_i, \tilde{x}_0), \ddot{\theta}(t_i, \tilde{x}_0)) + 10^{-2}\varepsilon_i^2\right\}_{i=1}^5, \quad (12)$$

where the ε_i^k (for $k = 1, 2$ and $i = 1, \dots, 5$) are independent 3×1 standard normal vectors.

We are interested in learning the vector field f over the set $\Omega = [-\pi, \pi]^2$ from the training data in (12) and the side information below, which could be derived from contextual knowledge without knowing f. We parameterize our candidate vector field p 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 $p_2(\theta, \dot{\theta})$.

- **Sign symmetry (Sym).** The pendulum system in Figure 2 is obviously symmetric with respect to the vertical dotted axis. Then, our candidate vector field p needs to satisfy the same symmetries.
$$p(-\theta, -\dot{\theta}) = -p(\theta, \dot{\theta}) \quad \forall(\theta, \dot{\theta}) \in \Omega.$$
Note that this is an affine constraint in the coefficients of the polynomial p.

- **Coordinate nonnegativity (Pos).** The only external force applied on the pendulum system is that of gravity; see Figure 2. This force pulls the 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 (8)). (Note that, because of the previous symmetry side information, we actually only need to impose the first of these two constraints.)

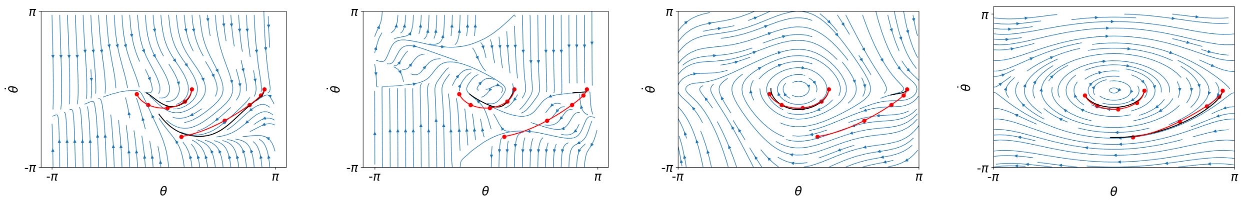
- **Hamiltonian (Ham).** The system in (11) is Hamiltonian. Indeed, in the simple pendulum model, there is no dissipation of energy (through friction for example), so the total energy
$$E(\theta, \dot{\theta}) = \frac{m}{2}\dot{\theta}^2 + \frac{1}{2}\frac{g}{l}(1 - \cos(\theta)) \quad (13)$$
is conserved. This energy is a Hamiltonian associated with the system. The two terms appearing in this equation can be interpreted physically as the kinetic and the potential energy of the system. Note that neither the vector field in (11) describing the dynamics of the simple pendulum nor the associated Hamiltonian in (13) 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 $p_1(\theta, \dot{\theta}) = -\frac{\partial H}{\partial \dot{\theta}}(\theta, \dot{\theta})$, and $p_2(\theta, \dot{\theta}) = \frac{\partial H}{\partial \theta}(\theta, \dot{\theta})$, but not the exact form of this Hamiltonian in (13). 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 for example on the coefficients of p.

Observe from Figure 3 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 SDP in Figure 3d is almost identical to the true dynamics in Figure 2 even though it is obtained only from 10 noisy samples on two trajectories. Figure 4 shows the benefit of adding side information even for predicting the future of a trajectory which is partially observed.

5. Approximation Results

In this section we present some density results for polynomial vector fields that obey side information. The proof of these results can be found in [1].

Theorem 1 *Fix a compact set $\Omega \subset \mathbb{R}^n$, a time horizon $T > 0$, and a desired accuracy $\varepsilon > 0$. Let $f : \Omega \to \mathbb{R}^n$ be a C^1 vector field that satisfies exactly one of the following side informa-*



(*a*) no side information (*b*) $\mathbf{Sym}$ (*c*) $\mathbf{Sym} \cap \mathbf{Pos}$ (*d*) $\mathbf{Sym} \cap \mathbf{Pos} \cap \mathbf{Ham}$

Figure 3: Streamplots of the polynomial vector fields of degree 5 returned by our SDPs for the simple pendulum as more side information constraints are added. In each case, the trajectories of the learned vector field starting from $(\frac{\pi}{4}, 0)^T$ and $(\frac{9\pi}{10}, 0)^T$ are plotted in black.

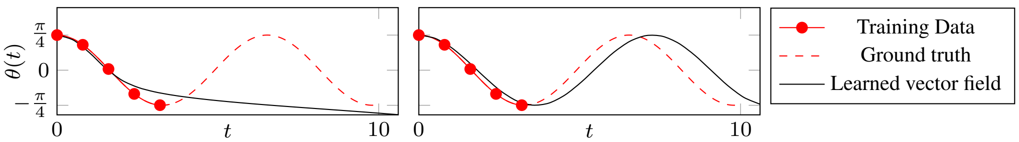


Figure 4: Comparison of the trajectory of the simple pendulum in (11) (dotted) starting from $(\frac{\pi}{4}, 0)^T$ with the trajectory from the same initial condition of the least-squares solution (left) and the vector field obtained from $\mathbf{Sym} \cap \mathbf{Pos} \cap \mathbf{Ham}$ (right).

tion constraints (see Section 2.1) (i) $\mathbf{Interp}(\{x_i, y_i)_{i=1}^m\}$*, (ii)* $\mathbf{Sym}(A, B)$*, (iii)* $\mathbf{Pos}(\{\trianglerighteq_i, B_i\}_{i=1}^n)$*, (iv)* $\mathbf{Mon}(\{\trianglerighteq_{i,j}, B_{i,j}\}_{i,j=1}^n)$*, (v)* $\mathbf{Inv}(B)$*, where* $B = \{x \in \mathbb{R}^n \mid h_i(x) \geq 0, \ i = 1, \ldots, m\}$ *for some* C^1 *concave functions* h_i *that satisfy* $h_i(x_0) > 0, \ i = 1, \ldots, m$*, for some* $x_0 \in \Omega$*, (vi)* $\mathbf{Grad}$ *or* $\mathbf{Ham}$*. Then there exists a polynomial vector field* $p : \mathbb{R}^n \to \mathbb{R}^n$ *such that* p *satisfies the same side information as* f*, and the trajectories of* p *and* f *starting from the same initial condition together with their first time derivatives remain within* ε *for all time* $t \in [0, T]$*.*

A natural question is whether the previous theorem could be generalized to allow for polynomial approximation of vector fields satisfying *combinations* of side information. It turns out that the answer is negative in general [1]. For this reason, we introduce the following notion of approximately satisfying side information.

Definition 1 *(δ-satisfiability) For any* $\delta > 0$ *and any side information* S *presented in Section 2.1, we say that a vector field* f δ*-satisfies* S *if for any equality constraint* $a = b$ *(resp. inequality constraint* $a \leq b$*) appearing in the definition of* S*, the vector field* f *satisfies the modified version* $|a - b| \leq \delta$ *(resp.* $a \leq b + \delta$*).*

Example 2 *A vector field* f δ*-satisfies the side information* $\mathbf{Interp}(\{x_i, y_i\}_{i=1}^m)$ *if* $\|f(x_i) - y_i\| \leq \delta$ *for* $i = 1, \ldots, m$*, and* δ*-satisfies the side information* $\mathbf{Pos}(\{\geq, B_i\}_{i=1}^n)$ *if* $f_i(x) \geq -\delta \ \forall x \in B_i$ *for* $i = 1, \ldots, n$*.*

The assumption of δ-satisfiability is reasonable because most optimization solvers return an approximate solution anyway. The following theorem shows that polynomial vector fields can approximate any vector field f and satisfy the same side information as f (up to an arbitrarily small error tolerance δ).

Theorem 2 *Fix a compact set* $\Omega \subset \mathbb{R}^n$*, a time horizon* $T > 0$*, a desired accuracy* $\varepsilon > 0$*, and a tolerance for error* δ*. Let* $f : \Omega \to \mathbb{R}^n$ *be a* C^1 *vector field that satisfies any combination of the six side information presented in Section 2.1. Then there exists a polynomial vector field* $p : \mathbb{R}^n \to \mathbb{R}^n$ *such that the trajectories of* p *and* f *starting from the same initial condition together with their first time derivatives remain within* ε *for all time* $t \in [0, T]$*, and* p δ*-satisfies the same combination of side information as* f*. Moreover,* δ*-satisfiability of side information comes with a sum of squares certificate of the form in (8).*

Acknowledgments

The authors are grateful to Charles Fefferman, Georgina Hall, Frederick Leve, Clancey Rowley, Vikas Sindhwani, and Ufuk Topcu for insightful questions and comments.

References

[1] Amir Ali Ahmadi and Bachir El Khadir. Learning dynamical systems with side information. *In preparation*, 2020.

[2] Mohamadreza Ahmadi, Ufuk Topcu, and Clarence Rowley. Control-oriented learning of Lagrangian and Hamiltonian systems. In *Annual American Control Conference*, pages 520–525, 2018.

[3] Roy M. Anderson, B. Anderson, and Robert M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[4] Ya-Chien Chang, Nima Roohi, and Sicun Gao. Neural Lyapunov control. In *Advances in Neural Information Processing Systems*, pages 3240–3249, 2019.

[5] Ching-An Cheng and Han-Pang Huang. Learn the Lagrangian: A vector-valued RKHS approach to identifying Lagrangian systems. *IEEE Transactions on Cybernetics*, 46(12):3247–3258, 2015.

[6] Mihaela Curmei and Georgina Hall. Nonnegative polynomials and shape-constrained regression. *In preparation*, 2020.

[7] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, pages 3240–3249, 2019.

[8] Georgina Hall. *Optimization over nonnegative and convex polynomials with and without semidefinite programming*. PhD thesis, Princeton University, 2018.

[9] J William Helton and Jiawang Nie. Semidefinite representation of convex sets. *Mathematical Programming*, 122(1):21–64, 2010.

[10] Hassan K Khalil. *Nonlinear Systems*. Prentice-Hall, 2002.

[11] Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[12] Jean B Lasserre. Convexity in semialgebraic geometry and polynomial optimization. *SIAM Journal on Optimization*, 19(4):1995–2014, 2009.

[13] Franz Lukács. Verschärfung des ersten Mittelwertsatzes der Integralrechnung für rationale Polynome. *Mathematische Zeitschrift*, 2(3):295–305, 1918.

[14] Pablo A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, May 2000.

[15] Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2, Ser. B):293–320, 2003.

[16] Mihai Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42(3):969–984, 1993.

[17] Hayden Schaeffer, Giang Tran, Rachel Ward, and Linan Zhang. Extracting structured dynamical systems using sparse optimization with very few samples. *arXiv preprint arXiv:1805.04158*, 2018.

[18] Vikas Sindhwani, Stephen Tu, and Mohi Khansari. Learning contracting vector fields for stable imitation learning. *arXiv preprint arXiv:1804.04878*, 2018.

[19] Sumeet Singh, Vikas Sindhwani, Jean-Jacques Slotine, and Marco Pavone. Learning stabilizable dynamical systems via control contraction metrics. In *Workshop on Algorithmic Foundations of Robotics*, 2018.