Nonlinear System Identification With Prior Knowledge on the Region of Attraction

Mohammad Khosravi, Graduate Student Member, IEEE, and Roy S. Smith, Fellow, IEEE

Abstract—We consider the problem of nonlinear system identification when prior knowledge is available on the region of attraction (ROA) of an equilibrium point. We propose an identification method in the form of an optimization problem, minimizing the fitting error and guaranteeing the desired stability property. The problem is approached by joint identification of the dynamics and a Lyapunov function verifying the stability property. In this setting, the hypothesis set is a reproducing kernel Hilbert space, and with respect to each point of the given subset of the ROA, the Lie derivative inequality of the Lyapunov function imposes a constraint. The problem is a non-convex infinite-dimensional optimization with an infinite number of constraints. To obtain a tractable formulation, only a suitably designed finite subset of the constraints are considered. The resulting problem admits a solution in form of a linear combination of the sections of the kernel and its derivatives. An equivalent finite dimension optimization problem with a quadratic cost function subject to linear and bilinear constraints is derived. A suitable change of variable gives a convex reformulation of the problem. The method is demonstrated by several examples.

Index Terms—Nonlinear system identification, region of attraction, prior knowledge, convex optimization.

I. INTRODUCTION

Nonlinear system identification has received significant attention due to its potential in modeling various phenomena in science and engineering [1]. Given the measurement data, the techniques of optimization, statistics, and system identification are used to mathematically model physical systems. In many situations modeling involves more than fitting nonlinear dynamics to the measurement data; one should include as prior knowledge additional features that are expected according to our understanding of the system. For example, in [2], [3] stabilizability of the dynamics is considered as a part of the identification problem.

Identification of stable continuous-time dynamical systems has been studied in [4]–[9] mainly motivated by imitation learning. In [5], Gaussian mixtures are used for modeling the dynamics with guaranteed global stability. In [6], a two-stage approach is presented where, first a parametric Lyapunov function as well as a model for the dynamics are learned, and then, the learned dynamics is stabilized using the Lyapunov function. The approach presented in [8] models the dynamics as a weakly nonlinear stable time-varying system which consists of a stable linear part for capturing the baseline behavior, and a nonlinear part to account for more complex phenomena, and a phase variable for the coupling these two parts. An identification method is introduced for learning a globally stable system in [9], however, the stability condition is only imposed locally on the data points by forcing the eigenvalues of Jacobian to be negative at sampling points. Similar to the current work, the hypothesis space in [9] is a smooth vector-valued reproducing kernel Hilbert space (SVRKHS) [10], [11].

We propose here a nonlinear system identification method designed to include the potentially available knowledge on a subset of the region of attraction (ROA) of a stable equilibrium point. This information might be attained by inspecting the measurement data, performing suitable experiments, using data-driven techniques for learning the ROA, or simply deduced from the physics of the system as can easily be done for pendulums and mass-spring-damper modeled systems [12]–[14]. Assuming that this stability property can be verified by a quadratic Lyapunov function, the problem is formulated as the joint estimation of the dynamics and the positive definite matrices determining the Lyapunov function. The resulting formulation is a non-convex optimization problem over an infinite dimensional space with infinite number of bilinear constraints, arising from the Lie derivative of the Lyapunov function with respect the points of the given subset of ROA. In order to make the problem tractable, we first introduce a suitable finite subset of the given subset of ROA such that verifying the Lie derivative inequality on these points guarantees the desired stability property. We derive an equivalent formulation as an optimization problem with a quadratic cost function, and linear and bilinear constraints. This problem admits a solution with a linear parametric representation in a special structure. Using a non-obvious change of variables, we derive a finite dimension convex reformulation of the problem. The method is demonstrated numerically by means of two examples.
II. Notations and Preliminaries

The set of natural numbers, the set of non-negative integers, the set of real numbers, the n-dimensional Euclidean space and the space of n by m real matrices are denoted by $\mathbb{N}$, $\mathbb{Z}_+\cup\{0\}$, $\mathbb{R}$, $\mathbb{R}^n$ and $\mathbb{R}^{n \times m}$ respectively. The identity matrix and zero vector in the Euclidean space are denoted by $I$ and $\mathbf{0}$ respectively. The set of positive definite (resp. positive semidefinite) matrices in $\mathbb{R}^{n \times n}$ is denoted by $\mathbb{S}^+_{nn}$ (resp. $\mathbb{S}^0_{nn}$). For any pair of symmetric matrices $X, Y \in \mathbb{R}^{n \times n}$, we write $X \succeq Y$ if $X - Y \in \mathbb{S}^+_{nn}$. Given $W \in \mathbb{S}^+_{nn}$, $\| \cdot \|_W$ is a norm on $\mathbb{R}^n$ defined as $\|x\|_W := x^T W x$, for any $x \in \mathbb{R}^n$. When $W = I$, we drop the subscript $W$.

Similarly, for any matrix $A \in \mathbb{R}^{n \times m}$, the norm of $A$ is denoted by $\|A\|$ and defined as $\|A\| := \sup_{x \in \mathbb{R}^n, \|x\|_1 \leq 1} \|Ax\|$. The (hyper)ball in $\mathbb{R}^n$ with center $c$ and radius $r > 0$ is denoted by $B(c, r)$ and defined as $B(c, r) := \{x \in \mathbb{R}^n | \|x - c\|_2 < r\}$. For a vector $\alpha = (\alpha_i)_{i=1}^n \in \mathbb{Z}^+_n$, define $|\alpha| := \sum_{i=1}^n \alpha_i$. For function $f : \mathbb{R}^n \to \mathbb{R}$, we denote the partial derivative with respect to the $i$th argument by $\frac{\partial f}{\partial x_i}$. If $f : \mathbb{R}^n \to \mathbb{R}^m$, then $\frac{\partial f}{\partial x}(x)$ is denoted as $[\frac{\partial f_1}{\partial x_1}(x), \ldots, \frac{\partial f_m}{\partial x_m}(x)]^T$ where $f_i$ is the $i$th coordinate of $f$.

The derivative operator is denoted by $D$, i.e., $Df$ is the derivative of $f$. The interior of set $\mathcal{X}$ is denoted by $\mathring{\mathcal{X}}$. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^n$ such that $\mathcal{X}$ is the closure of $\mathring{\mathcal{X}}$. Then, $C^2(\mathcal{X}, \mathbb{R}^n)$ is defined as the space of functions $g : \mathcal{X} \to \mathbb{R}^n$ where $\partial^2 g|_{\mathcal{X}}$ is well-defined and has a continuous extension to $\mathcal{X}$, for any $\alpha \in \mathbb{Z}^+_n$ with $|\alpha| \leq s$. Define $C^2(\mathcal{X} \times \mathcal{X}, \mathbb{R}^{n \times n})$ similarly. Let $\mathcal{Y}$ be a set and $\mathcal{C} \subseteq \mathcal{Y}$. We define the function $\mathcal{I}_C$ as $\mathcal{I}_C(y) = 0$, if $y \in \mathcal{C}$ and $\mathcal{I}_C(y) = \infty$, otherwise.

III. Problem Statement

Let $\mathcal{U}$ be an open domain in $\mathbb{R}^n$ and $f : \mathcal{U} \to \mathbb{R}^n$ be an unknown vector field defined on $\mathcal{U}$ which is $C^2(\mathcal{U}, \mathbb{R}^n)$. Consider the corresponding dynamical system defined as

$$\dot{x} = f(x), \quad x(0) = x_0,$$

where $x_0 \in \mathcal{U}$ is the initial point. Denote the solution of (1), at time instant $t \geq 0$, by $x(t; x_0)$. Let the origin be an asymptotically stable equilibrium of dynamical system (1). Also, let the corresponding region of attraction (ROA) be denoted by $\Omega_{\text{ROA}}$, i.e., we have $\Omega_{\text{ROA}} := \{x_0 \in \mathcal{U} | \lim_{t \to \infty} x(t; x_0) = \mathbf{0}\}$. Let $\Omega$ be a known inner compact approximation for the region of attraction of the origin $\Omega_{\text{ROA}}$.

Consider a set of trajectories of system, $\{x(\cdot; x_0^l) | 1 \leq l \leq n_T\}$, where $x_0^1, \ldots, x_0^{n_T} \in \mathcal{U}$ are the corresponding initial points. For any $i = 1, \ldots, n_T$, suppose that the $i$th trajectory is sampled at time instants $0 \leq t_1^i < t_2^i < \cdots < t_n^i$, where $n_i \in \mathbb{N}$. Let $x_k^i$ denote measured $x(t_k^i; x_0^l)$, for $1 \leq k \leq n_i$. The measurements are given to a preprocessing unit for estimating the time derivative of $x(\cdot; x_0^l)$ at the sampling time instants. This unit might use techniques such as finite-difference procedures, nonlinear regression methods together with analytical or numerical derivative calculation, or robust derivative estimation methods [15]. Let these estimations be denoted by $\hat{x}_k^l$, for $1 \leq k \leq n_i$. One should note that $\hat{x}_k^l$ is an approximation of $f(x_k^l)$. Considering these samples of trajectories and their estimated derivatives, we get a set of data, denoted by $\mathcal{D}$, which contains $\{(x_k^l, \hat{y}_k^l)\}$ pairs. For notation simplicity, we drop the superscripts and simply show set $\mathcal{D}$ as $\{(x_j, y_j) | 1 \leq j \leq m\}$, where $m = \sum_{i=1}^{n_T} n_i$.

**Problem:** Given that origin is a stable equilibrium point of (1) and the set $\Omega$ is provided as the prior knowledge about the region of attraction of the origin, the problem is to estimate the unknown vector field $f$, in a given class of functions $f \in \mathcal{F} \subseteq C^2(\mathcal{U}, \mathbb{R}^n)$, using the set of data $\mathcal{D}$.

In the next section, we introduce a tractable formulation of this problem as a nonparametric estimation.

IV. Main Results: Identification Method

We know that $f$ satisfies the constraint that $x = \mathbf{0}$ is an equilibrium point of (1), i.e., $f(\mathbf{0}) = \mathbf{0}$. Moreover, we know that $x = \mathbf{0}$ is stable and $\Omega$ is a subset of the corresponding region of attraction. Assume that these stability features of $f$ can be verified by an unspecified quadratic Lyapunov function $V(x) = \frac{1}{2} x^TPx$ where here $P$ is a positive definite matrix. More precisely, there exist an unknown $\epsilon > 0$ and an unknown positive definite matrix $P \succeq I$ such that

$$DV(x)/f(x) = x^TPx \leq -\epsilon \|x\|^2, \quad \forall x \in \Omega.$$  

(2)

In the estimation problem, we need to minimize the fitting error, $\sum_{i=1}^{n_T} \|y_i - f(x_i)\|^2_{\mathbf{W}}$, subject to $f \in \mathcal{F}$, $f(\mathbf{0}) = \mathbf{0}$ and (2). There are two main issues: the correct choice of function class $\mathcal{F}$, and dealing with the (uncountable) infinite number of constraints in (2). These issues are addressed in this section.

A. From Infinite to Finite Number of Constraints

To develop a tractable optimization problem we need to introduce a finite number of constraints implying stability condition (2). To this end, the notion of $(\alpha, \beta)$-grid is introduced in the next definition.

**Definition 1:** Let $\mathcal{Z} := \{z_1, \ldots, z_{n_g}\}$ be a finite subset of $\Omega\setminus\{\mathbf{0}\}$. We say $\mathcal{Z}$ is an $(\alpha, \beta)$-grid or a grid for $\Omega$ if

$$\Omega \subseteq \bigcup_{i=1}^{n_g} B(z_i, \alpha\|z_i\|) \cup B(\mathbf{0}, \beta).$$

(3)

Next theorem shows the role of $(\alpha, \beta)$-grid in the estimation problem. Let $L_1, L_2$ be constants such that

$$L_1 \geq \sup_{x \in \Omega} \|Df(x)\|,$$

$$L_2 \geq \sup_{x \in B(h_1, h_2) \cup B(0, 1)} \|D^2f(x)(h_1, h_2)\|.$$  

(4)

Since $f$ is $C^2(\mathcal{U}, \mathbb{R}^n)$, there exist $L_1, L_2 < \infty$. Indeed, $L_1$ and $L_2$ are the Lipschitz constant for $Df$ and $D^2f$, respectively.

**Theorem 1:** Let $\mathcal{Z} := \{z_1, \ldots, z_{n_g}\} \subset \Omega\setminus\{\mathbf{0}\}$ and $P \in \mathbb{S}^+_{nn}$ be such that $P \succeq I$ and

$$z_i^T Pz_i \geq -\|z_i\|^2, \quad 1 \leq i \leq n_g,$$

$$\frac{1}{2} (P Df(\mathbf{0}) + Df(\mathbf{0})^T P) \succeq -I.$$  

(5)

Given $\epsilon \in (0, 1)$, let $\alpha$ and $\beta$ be real positive scalars where

$$\alpha \leq \left(1 + \frac{L_1\|P\|}{\epsilon + L_1\|P\|}\right)^{-\frac{1}{2}} - 1, \quad \beta \leq \frac{1}{L_2\|P\|}$$

(6)

and $\mathcal{Z}$ be an $(\alpha, \beta)$-grid for $\Omega$. Then, the following holds

$$x_i^T P x_i \leq -\epsilon \|x_i\|^2, \quad \forall x_i \in \Omega.$$  

(7)
Proof: Let $x \in \Omega$. Since $Z$ is an $(\alpha, \beta)$-grid, then by (3), we know that $x$ belongs to $B(0, \beta)$, or it belongs to $B(z_i, \alpha|z_i|)$, for some $1 \leq i \leq n_g$.

Case I: Assume that $x \in B(0, \beta)$, i.e., $\|x\| < \beta$. Since, $f$ is $C^2(U, \mathbb{R}^{n_0})$ and $f(0) = 0$, from a Taylor expansion at the origin, we have that $f(x) = Df(0)x + r(x)$, where $r : U \to \mathbb{R}^n$ is a $C^2(U, \mathbb{R}^n)$ function such that sup$_{x \in \Omega, x \neq 0} \|r(x)\| \leq L_2$.

Accordingly, one can easily see that $x^Tf(x) = 1/2x^TPDf(0)x + Df(0)x^TPx + r^Tx$, subsequently, Due to (5) and the Cauchy-Schwartz inequality, we have $x^Tf(x) \leq -\|x\|^2 + \|Df(0)x\|^2$. According to (6) and since $\|x\| < \beta$, we have $x^Tf(x) \leq -\|x\|^2$.

Case II: Assume that $x \in B(z_i, \alpha|z_i|)$, i.e., $x = z_i + a$ where $a$ is a vector such that $\|a\| < \alpha|z_i|$. From the triangle inequality, we have that

$$-\epsilon|z_i|^2(1 + \alpha)^2 \leq -\epsilon(\|z_i\|^2 + \|a\|^2) \leq -\epsilon\|x\|^2.$$  

Due to (4), $L_1$ is a Lipschitz constant for $f$. This implies $\|f(z_i + a) - f(z_i)| \leq L_1(\|a\|)$ and $\|f(z_i + a)\| \leq L_1(\|z_i\| + \|a\|)$, due to $f(0) = 0$ and the triangle inequality. Subsequently, from $\|Pz_i\| \leq \|P\|\|z_i\|$, $\|Pa\| \leq \|P\|\|a\|$, and the Cauchy-Schwartz inequality, we have

$$x^Tf(x) = z_i^TPf(z_i) + a^TPf(z_i + a) - f(z_i) + a^TPf(z_i + a) \leq -\|z_i\|^2 + \|z_i\|\|P\|L_1\|a\| + \|a\|\|P\|L_1(\|z_i\| + \|a\|).$$

Since $\|a\| < \alpha|z_i|$, one can conclude that

$$x^Tf(x) < -\|z_i\|^2(1 - 2\|P\|L_1\alpha - \|P\|L_1\|a\|)^2 \geq \epsilon(1 + \alpha)^2.$$  

Accordingly, due to (8) and (9), we have that $x^Tf(x) < -\epsilon(1 + \alpha)^2\|x\|^2 \leq -\epsilon\|x\|^2$. This concludes the proof.

Theorem 2: For any $\alpha, \beta > 0$, an $(\alpha, \beta)$-grid exists for $\Omega$.

Proof: Since $\Omega \subset \bigcup_{z \in \Omega, z \neq 0} B(z, \alpha|z|) \cup B(z, \beta)$ and $\Omega$ is a compact set, this open cover has a finite sub-cover. Hence, there exist $\{z_1, \ldots, z_{n_g}\} \subset \Omega \setminus \{0\}$ such that (3) holds.

Remark 1: Theorem 1 guarantees that for satisfying the infinite number of constraints (2), it is enough to satisfy the finite number of constraints (5), given a suitable grid for $\Omega$. The grid may be designed based on an ad hoc approaches, especially when $\Omega$ has a specific shape like a hyperball, or by generating a random finite subset of $\Omega$ with suitable size.

B. Approximating the Dynamics in Smooth Vector-Valued Reproducing Kernel Hilbert Spaces

The function class taken for approximating the unknown vector field is a type of Hilbert spaces called smooth vector valued reproducing kernel Hilbert spaces (SVKHS) which are introduced below (see [3], [10], [11] for more details). Based on the structure of SVKHS, we prove that the problem admits a solution with a specific finite linear parametric form. This reduces the optimization problem to the coefficients of this representation and a tractable finite-dimensional problem results.

Let $X$ be a compact subset of $\mathbb{R}^n$ with non-empty interior $X^o$ such that $X$ is the closure of $X^o$ [10] and $\Omega \subset X^o$.

Definition 3: A Smooth Vector-valued Reproducing Kernel Hilbert Space (SVKHS), denoted by $H$, is a Hilbert space of functions $g \in C^0(X, \mathbb{R}^n)$ such that for any $x \in \Omega$, we have sup$_{x \in H, \|x\| \leq 1} \|g(x)\| < \infty$.

Definition 4: The function $K \in C^2(X \times X, \mathbb{R}^{n \times n})$ is an operator-valued positive-definite Mercer kernel [3] when for any $m \in \mathbb{N}$, $x, y, x_1, \ldots, x_m \in X$ and $a_1, \ldots, a_m \in \mathbb{R}^n$, we have $K(x, y) = E(x, y)^T$ and $\sum_{i, j} a_i^T K(x_i, x_j) a_j = 0$.

For any $x \in X$, let $K_x$ denote the function defined by $K_x(., x) : X \to \mathbb{R}^n$. This is called the section of kernel $K$ at $x$ or the feature map.

Theorem 3 [3]: With respect to any Mercer kernel $K \in C^2(X \times X, \mathbb{R}^{n \times n})$, there exists a SVKHS of functions $g \in C^0(X, \mathbb{R}^n)$, denoted by $H_K$, and endowed with inner product $(\cdot, \cdot)_H_K$ and norm $\|\cdot\|_H_K$, such that for any $(x, y) \in X \times X$ and for any $\alpha \in \mathbb{R}^n$, we have $i) \delta_{a}K_x y \in H_K$, and $ii) (g, \delta_{a}K_x y)_{H_K} = y^T a g$, for all $g \in H_K$. The second feature is called the reproducing property.

We assume that the kernel is suitably chosen such that function $h : X \to \mathbb{R}^n$, defined as $h(x) = x$, belongs to $H_K$. A simple example is $K(x, y) = \delta(x, y)$, where $\delta$ is a polynomial kernel. Moreover, let $T \in S_{+}^{n \times n}$ be a positive-definite finite-dimensional transformation such that $Tf \in H_K$. In other words, this implies that $g \in H_K$, where $x : X \to \mathbb{R}^n$ is defined as $g(x) = T f(x)$, for any $x \in X$. For example $T$ might be a scaling of the identity matrix.

Let the fitting loss or the error function, denoted by $L_D$, be the function $L_D : H_K \to \mathbb{R}$ defined as $L_D(f) := \sum_{i=1}^{n_g} y_i - f(x_i)^2$, where $W \in S_{+}^{n \times n}$ is a weight matrix. Additionally, we can consider a suitable kernel-based regularization due to $Tf \in H_K$. More precisely, let the regularization function $\mathcal{R} : H_K \to \mathbb{R}_{\geq 0}$ be defined as $\mathcal{R}(Tf) := \|Tf\|_{H_K}^2$. The identification problem is now formulated as following

$$\min_{f \in C^2(U, \mathbb{R}^n), P \in S_{+}^{n \times n}} \sum_{i=1}^{n_g} \|y_i - f(x_i)^2\_W + \lambda\|Tf\|_{H_K}^2, \quad \mathcal{R}(Tf) \leq \|Tf\|_{H_K}^2,$$

where $\lambda > 0$ is the regularization weight and $C$ is the set of $C^2(U, \mathbb{R}^n)$ vector fields such that $Tf \in H_K$ and we have that $x = 0$ is a stable equilibrium point which is attracting in $\Omega$.

Due to Theorem 1, it is sufficient to take a suitable $(\alpha, \beta)$-grid $Z = \{z_1, \ldots, z_{n_g}\}$ and solve optimization problem over $Z$.

$$\min_{g \in H_K, P \geq 1} \sum_{i=1}^{n_g} \|y_i - T^{-1}g(x_i)^2\_W + \lambda\|g\|_{H_K}^2, \quad \mathcal{R}(Tf) \leq \|Tf\|_{H_K}^2, \quad \mathcal{R}(Tf) \leq \|Tf\|_{H_K}^2,$$  

s.t. $z_i^TP^{-1}g(z_i) \leq -\|z_i\|^2, \quad 1 \leq i \leq n_g,$

$$\frac{1}{2}(PDf(0) + Df(0)^TP) - \|z_i\|^2, \quad f(0) = 0,$$

$$Tf \in H_K,$$

$$P \geq 1.$$  

(11)

The existence of such grid is guaranteed by Theorem 1.  Rewriting optimization problem (11) in terms of $g$, one has

$$\min_{g \in H_K, P \geq 1} \sum_{i=1}^{n_g} \|y_i - T^{-1}g(x_i)^2\_W + \lambda\|g\|_{H_K}^2, \quad \mathcal{R}(Tf) \leq \|Tf\|_{H_K}^2, \quad \mathcal{R}(Tf) \leq \|Tf\|_{H_K}^2,$$  

s.t. $z_i^TP^{-1}g(z_i) \leq -\|z_i\|^2, \quad 1 \leq i \leq n_g,$

$$\frac{1}{2}(PT^{-1}Dg(0) + Dg(0)^TP) - \|z_i\|^2, \quad g(0) = 0.$$  

(12)

The problem (12) is a non-convex infinite-dimensional optimization and therefore, it is not tractable. However, in
order to address this issue, we derive a finite dimensional problem equivalent to (12).

With respect to a given $P \in S^n_{++}$, we define $F_P$ as

$$
F_P := \left\{ g \in \mathbb{H}_K | z_i^2(P + T)^{-1}g(z_i) \leq -\|z_i\|^2, \quad 1 \leq i \leq n, \right\}
$$

$$(13)$$

Theorem 4: For any $P \succeq I$, the set $F_P$ is a non-empty, closed and convex subset of $\mathbb{H}_K$.

Proof: Since $P, T \in S^n_{++}$, the eigenvalues of matrix $M := \frac{1}{2}(PT + T^TP)$ are positive. Define function $g_P : X \rightarrow \mathbb{R}^n$ as $g_P(x) = -2y^T x$ where $y \in \mathbb{R}^{n_0}$ is smaller than the eigenvalues of $M$. Since $z_i^2(P + T)^{-1}g(z_i) \leq -\|z_i\|^2$, and $\frac{1}{2}(PT + T^TP) = -2y^T x \leq -I_P$. Hence, $g_P(x) = 0$. Therefore, $g_P \in F_P$ and $F_P$ is not empty.

The convexity of $F_P$ is due to the linear dependency of the left-hand sides of the constraints with respect to $g$. Now, let $(g_k)_{k=1}^\infty$ be a sequence converging to $g \in \mathbb{H}_K$. Let $y$ be an arbitrary vector in $\mathbb{R}^n$. Then, for any $x \in X$ and any $\alpha \in \mathbb{R}_{+}$, due to the reproducing property and the Cauchy-Schwarz inequality, we have $|y^T \langle \alpha g(x) - \alpha g(x) \rangle| = |\|y\|^2 - \|g(x)\|^2| \leq \|g(x)\|^2$. Therefore, as $\alpha \rightarrow \infty$, we have $\|g(x)\|^2 \rightarrow 0$, and subsequently, $g(x) \rightarrow 0$. Since for all $k \in \mathbb{N}$, $g_k$ satisfies the constraints and due to linear dependency of constraints on $g_k$ and $x$, they are also satisfied by $g$ and $g \in F_P$. Hence, $F_P$ is a closed subset of $\mathbb{H}_K$.

For ease of notation, we define $P := n + n_g, \quad m := 1 + p + n, \quad I_p := \{0, \ldots, n + n_g\}, \quad I_p := \{1 + n, \ldots, p\}, \quad I_n := \{0, \ldots, m - 1\}$, and also, set $x_0 = 0$, and $x_i = x_{i-1} - x_i$, for $i \in I_p$.

Theorem 5: For any $P \succeq I$ and $\lambda > 0$, the optimization problem

$$
\min_{g \in F_P} \sum_{i=1}^p \|y_i - T^{-1}g(x_i)\|_W^2 + \|g\|_{\mathbb{H}_K}^2
$$

has a unique solution, denoted by $g_P^*$. Moreover, there exist vectors $(a_i)^{n-1}_{i=1}$ such that, $g_P^*$, the solution of (14), is in the following form

$$
g_P^* = \sum_{i=0}^p K_n a_i + \sum_{j=1}^n a_j \partial K_0 a_{j+p}.
$$

Proof: Define $J : \mathbb{H}_K \rightarrow \mathbb{R}(I \cup (+\infty))$ as $J(g) := \sum_{i=1}^p \|y_i - T^{-1}g(x_i)\|_W^2 + \|g\|_{\mathbb{H}_K}^2 + I_F_P(g)$, for any $g \in \mathbb{H}_K$. According to Theorem 4, $F_P$ is a non-empty, closed, and convex set, and therefore $J_P$ is a proper lower-semicontinuous convex function [16]. Let $g_P$ be the element of $\mathbb{H}_K$ introduced in the proof of Theorem 4. Since $g_P \in F_P$, we have $J_P(g_P) = 0$. Also, we know that $\sum_{i=1}^p \|y_i - T^{-1}g(x_i)\|_W^2 < \infty$. Therefore, $\sum_{i=1}^p \|y_i - T^{-1}g(x_i)\|_W^2$ is a proper and continuous convex function of $g$. Since $\lambda > 0$ and $\|g\|_{\mathbb{H}_K} < \infty$, we know that $J$ is proper, lower-semicontinuous and strongly convex. Therefore, $\min_{g \in \mathbb{H}_K} J(g)$ has a unique (finite) solution [16], which means that (14) admits a unique solution with finite cost. Define set $V \subseteq \mathbb{H}_K$ as $V := \{ \sum_{i=0}^p K_n a_i + \sum_{j=1}^n a_j \partial K_0 a_{j+p} | a_0, \ldots, a_m \in \mathbb{R}^n \}$. This is a finite-dimensional subspace of $\mathbb{H}_K$ and consequently, it is a closed subspace. Hence, one can decompose $g_P^* = g_P + g_1$, where $g_0 \in V$ and $g_1 \in V$. Therefore, for any $y \in \mathbb{R}^n$ and for any $i \in I_p$, from the reproducing property, we have $y^T g_1(x_i) = \langle g_0, K_0 \rangle_\mathbb{H}_K = 0$ and subsequently, we have $y^T g_1(x_i) = y^T g_0(x_i)$. Accordingly, one can conclude that $g(x_i) = g_0(x_i)$, for any $i \in I_p$. Similarly, due to the reproducing property, we have $y^T Dg_0(x_0) = [y^T \partial_1 K_0(0) \ldots, y^T \partial m K_0(0)] = [g_{01}, \partial_1 K_0(0), \ldots, g_{0m}, \partial m K_0(0)] = 0_T$, which shows that $Dg_0(0)$ is zero and subsequently, $Dg_1(0) = 0_T$. Since $g_P^* \in F_P$, $Dg_1(0) = 0_T$ and $g(x_i) = g_0(x_i)$, for any $i \in I_p$, one can see that $g_1$ satisfies the constraints defining $F_P$, i.e., $g_1 \in F_P$ and $J_P(g_1) = 0$. Similarly, we have $\sum_{i=1}^n \|y_i - T^{-1}g_1(x_i)\|^2_W = \sum_{i=1}^n \|y_i - T^{-1}g_0(x_i)\|^2_W$ and $\|g_0\|^2_{\mathbb{H}_K} = \|g_1\|^2_{\mathbb{H}_K} + \|g_0\|^2_{\mathbb{H}_K}$. Therefore, $g_0 = 0$, otherwise $g_1$ is a feasible solution with objective value smaller than minimum of the objective function. Hence $g_P^* = g_1$ and since we have $g \in \mathbb{H}_K$, $g_P^*$ has the form given in (15).

Without loss of generality, we take $K$ a diagonal kernel in the form of $K(x, y) = k(x, y)I$, where $K \in \mathbb{C}^{2 \times (X \times X, \mathbb{R})}$ is a scalar valued Mercer kernel. Note that the kernel $K$ is characterized by a number of constants called hyperparameters where here for the sake of more transparent discussion, we have dropped this dependency in the notation. The hyperparameters are required to be estimated which is commonly done by a cross-validation procedure. The hyperparameter estimation is a computationally demanding procedure, especially when the kernel has a large number of hyperparameters. Accordingly, $K$ being a diagonal kernel, significantly reduces the computational load of the final estimation problem.

The next theorem provides the finite dimensional version of (12). First, we need to introduce some preliminaries. Define $A := [a_0, \ldots, a_{m-1}] \in \mathbb{R}^{m \times n}$ and function $k : X \rightarrow \mathbb{R}^n$ as

$$
k(x) = [K_0(x), \ldots, K_m(x), \partial_1 K_0(x), \ldots, \partial m K_0(x)]^T,
$$

for $x \in X$. One can see that if $g$ is in the form of

$$
g = \sum_{i=0}^p \|K_n a_i + \sum_{j=1}^n a_j \partial K_0 a_{j+p},
$$

as in (15), we have $g = Ak$.

Lemma 1: Let $g$ be a function defined as $g := Ak$ where $A \in \mathbb{R}^{m \times n}$. Then, we have that $g \in \mathbb{H}_K$ and $\|g\|^2_{\mathbb{H}_K} = \text{tr}(AKA^T)$, where $K := [K_{ij}]_{i,j=0 \ldots m-1} \in \mathbb{R}^{m \times m}$ is a matrix defined element-wise as

$$
K_{ij} := \begin{cases} 
[K_0(x_i, x_j), & 0 \leq i, j \leq p, \\
\partial_p K_0(x_i, 0), & 0 \leq i \leq p < j < m, \\
\partial_p K_0(x_i, 0), & 0 \leq j \leq p < i < m, \\
\partial_{i-j} K_0(0, 0), & p + 1 \leq i, j \leq m.
\end{cases}
$$

Proof: Since $\text{tr}(AKA^T) = \sum_{i,j=0}^m a_i K_0 a_j$, from (16), and (18) and the reproducing property, the claim follows.

Define $K_i := k(x_i)$, for any $i \in I_m$ and $J \in \mathbb{R}^{m \times n}$ as the Jacobian of $k$ at $x = 0$, i.e., $J := [\partial_1 k(0), \ldots, \partial m k(0)]$. Hence, $g_P^* = g_1$ and since we have $g \in \mathbb{H}_K$, $g_P^*$ has the form given in (15).
Theorem 6: Consider the following optimization problem
\[
\begin{align*}
\min_{P \succeq I, A \in \mathbb{R}^{n \times n}} & \sum_{i=1}^{n} \| y_i - T^{-1} AK_i \|_{W}^2 + \lambda \text{tr}(AKAT) \\
\text{s.t.} & \sum_{i=1}^{n} \| y_i - T^{-1} g(x_i) \|_{W}^2 + \lambda \| g \|_{H_{\Omega}}^2.
\end{align*}
\] (19)

Then, (19) admits a solution. Moreover, with respect to each solution of (19), one can find a solution for (12).

Proof: The existence of solution is due to the fact that the cost in (19) is a quadratic function and the feasible set is non-empty and closed. Now, note that one can restate optimization problem (12) in the following equivalent form
\[
\begin{align*}
\min_{P \succeq I} \left( \min_{g \in \mathcal{F}_P} \sum_{i=1}^{n} \| y_i - T^{-1} g(x_i) \|_{W}^2 + \lambda \| g \|_{H_{\Omega}}^2 \right).
\end{align*}
\]

From Theorem 5, we know that the solution of the inner problem has a unique solution in the form of \( g(x) = A k(x) \), for a matrix \( A \in \mathbb{R}^{n \times n} \). Since \( D g(0) = A J \), due to Lemma 1 and by substituting \( g(x) = A k(x) \) in (12), optimization problem (19) results. Solving (19), we obtain \( A, P \), and subsequently, \( g^* = A k \) and \( f^* = T^{-1} g = T^{-1} A k \).

According to Theorem 6, for solving (12), it is enough to solve (19), which is a finite-dimensional non-convex optimization due to the bilinear constraints. This non-convexity can be a limiting issue especially when we have large number of variables. This issue is addressed in the following section.

C. Convex Reformulation

In the estimation problem (11), and subsequently (12), the matrices \( W \) and \( T \) are introduced as arbitrary positive definite matrices. One can see that the mathematical arguments (up to Theorem 6) only require the fact that \( W \) and \( T \) do not depend on \( g \). This provides the opportunity of choosing them such that a change of variables leads to a convex formulation. In fact, we set \( W := P^2 \) and \( T := P \). Since \( \| x \|_{P^2} = \| P x \|_2 \), for \( x \in \mathbb{R}^n \), we have \( \sum_{i=1}^{n} \| y_i - T^{-1} g(x_i) \|_{W}^2 \leq \sum_{i=1}^{n} \| P y_i - g(x_i) \|_2^2 \). Therefore, problem (12) can be modified to give
\[
\begin{align*}
\min_{g \in \mathcal{H}_{\Omega}, P \succeq I} & \sum_{i=1}^{n} \| P y_i - g(x_i) \|_{W}^2 + \lambda \| g \|_{H_{\Omega}}^2 \\
\text{s.t.} & \sum_{i=1}^{n} \| y_i - T^{-1} g(x_i) \|_{W}^2 + \lambda \| g \|_{H_{\Omega}}^2, \\
& 1 \leq i \leq n_g, \\
& \frac{1}{2} (D g(0) + D g(0)^\top) \preceq -I, \\
& g(0) = 0.
\end{align*}
\] (20)

One can see that (20) is a convex optimization problem which by Theorem 5 has a unique solution of the form \( g^* = A k \). By substituting this into (20), we obtain a finite problem as
\[
\begin{align*}
\min_{A \in \mathbb{R}^{n \times n}, P \succeq I} & \sum_{i=1}^{n} \| P y_i - AK_i \|_{W}^2 + \lambda \text{tr}(AKA^\top) \\
\text{s.t.} & \sum_{i=1}^{n} \| y_i - T^{-1} AK_i \|_{W}^2 + \lambda \text{tr}(AKA^\top), \\
& 1 \leq i \leq n_g, \\
& \frac{1}{2} (A J + J^\top A^\top) \preceq -I, \\
& AK_0 = 0.
\end{align*}
\] (21)

The identification procedure is summarized in Algorithm 1.

Algorithm 1 Nonlinear Identification With Prior Knowledge on ROA
1: Init: \( \mathcal{D} := \{ x_i, y_i \}_{i=1}^{n} \), \( \Omega \), kernel \( k \), \( L_1, L_2 \) and \( \epsilon \in (0, 1) \),
2: Generate \( (a, \beta) \)-grid \( \mathcal{Z} \) (see Definition 1 & Theorem 1),
3: Set \( x_0 = 0 \) and \( x_i+n_i = x_i \) for \( 1 \leq i \leq n_g \), and calculate vector \( K_i = k(x_i) \), for \( 1 \leq i \leq n_g \) (see (10)), matrix K (see (18)) and Jacobian matrix \( J := [\partial_1 k(0), \ldots, \partial_n k(0)] \),
4: Solve convex optimization (21),
5: Output: \( P, A \) and \( f \mathcal{I} \rightarrow \mathbb{R}^n \) defined as \( f = P^{-1} A k \).

V. NUMERICAL EXPERIMENTS

In this section, we discuss two numerical examples, one for illustration and one for comparison.

Example 1: Consider the dynamical system defined as
\[
\begin{align*}
\dot{x}_1 &= -4x_1 - 5x_2 - 6x_1^2 + x_1 x_2, \\
\dot{x}_2 &= -20x_1 - 4x_2 + 4x_1^2 x_2 + x_2^3.
\end{align*}
\] (22)

For system (22), \( x = 0 \) is a stable equilibrium point, which is attracting in \( \Omega := B(0, 1.5) \). Let this information be provided as prior knowledge. We consider two trajectories of the system, starting from \((-1, -1)\) and \((1, -1)\), and take samples from each of them at different 19 locations with an additive measurement noise. The calculated signal-to-noise-ratio (SNR) for \( (x_1)_{i=1}^{38} \) and \( (y_1)_{i=1}^{38} \) are respectively 30.6 dB and 26.5 dB. We take two approaches for identifying (22): 1) we utilize the prior knowledge on the ROA and solve (21), 2) only the stability of \( x = 0 \) is considered and we solve a modified version of (21) where \( P = I \) and the grid constraints are relaxed. Denote the corresponding solutions by \( \hat{f} \) and \( \tilde{f} \), respectively. The main difference in these approaches is the inclusion of the prior knowledge on ROA in the estimation method. The comparison of \( \hat{f} \) and \( \tilde{f} \) reflects the impact and the potential leverage of using the prior knowledge of the ROA on the estimation. We set \( \mathcal{Z} \) as \( \{1.5 \cdot 0.8(\cos(0.1\pi), \sin(0.1\pi))|1 \leq i \leq 15, 1 \leq j \leq 20\} \). Given these settings, we obtain estimations \( \hat{f} \) and \( \tilde{f} \). The calculated \( R \) squared in \( \Omega \) for \( \hat{f} \) and \( \tilde{f} \) is 91.4% and 73.7%, respectively. Now, let \( x_1, x_2 \) and \( x_3 \) denote the trajectories generated by \( \hat{f} \), \( \hat{f} \), and \( \tilde{f} \), respectively. We show three initial points \( x_0 = (0.8, 1.1) \), \( \hat{x}_2 = (1.3, -0.6) \) and \( \tilde{x}_2 = (-2, -0.5) \). The corresponding trajectories are shown in Fig. 1. For point \( x_0 = \hat{x}_0 \in \Omega \), all of the trajectories goes to the equilibrium point \( (0, 0) \). Trajectory \( x_j \) is closer to the trajectory of true system, \( x_j \), in comparison to \( x_j \). Starting from \( x_0 = \hat{x}_2 \in \Omega \), trajectories \( x_j \) and \( x_j \) stay close to each other and converge to \((0, 0)\), while \( x_j \) diverges. This confirms that the prior knowledge is satisfied by the estimated vector field \( \hat{f} \).

Finally, if \( x_0 = \hat{x}_2 \notin \text{ROA} \), trajectory \( x_j \) as well as trajectory \( x_j \) diverge, while \( x_j \) converges to \((0, 0)\).

Example 2: Consider the following dynamics
\[
\begin{align*}
\dot{x}_1 &= -x_1 + \frac{2x_1^2}{1 + x_2^2}, \dot{x}_3 = -2x_3 + \frac{x_1^2}{1 + x_2^2} - \frac{5}{1 + x_2^2}, \\
\dot{x}_2 &= -x_2 + \frac{x_1^2}{1 + x_1^2} - \frac{4}{1 + x_1^2}, \dot{x}_4 = -2x_4 + \frac{x_1^2}{1 + x_3^2},
\end{align*}
\]

which is an example of multistable biological system [17]. The system is initialized at 20 different points in \( x_i \) such that \( |x_i| \leq 0.55 \) and the corresponding trajectories are sampled with
VI. Conclusion

We have discussed nonlinear system identification when prior knowledge is available on a subset of the region of attraction (ROA) of an equilibrium point. The proposed method is a joint estimation of the dynamics and a Lyapunov function for the stability property. The problem is initially formulated as a bilinear infinite-dimensional optimization with infinite number of constraints, and then, an equivalent tractable convex reformulation is derived. A numerical example shows better performance with respect to noise than two existing methods. For future work, one may extend the proposed method for discrete-time dynamics. Another interesting direction is introducing an ADMM-like algorithm for the large scale situations.

References

[1] J. Schoukens and L. Ljung, “Nonlinear system identification: A user-oriented road map,” IEEE Control Syst. Mag., vol. 39, no. 6, pp. 28–99, Dec. 2019.
[2] Y. Sattar and S. Oymak, “Non-asymptotic and accurate learning of nonlinear dynamical systems,” 2020. [Online]. Available: arXiv:2002.08538.
[3] S. Singh, S. M. Richards, V. Sindhwani, J.-J. E. Slotine, and M. Pavone, “Learning stabilizable nonlinear dynamics with contraction-based regularization,” 2019. [Online]. Available: arXiv:1907.13122.
[4] S. Calinon, F. D’halluin, E. L. Sauser, D. G. Caldwell, and A. G. Billard, “Learning and reproduction of gestures by imitation,” IEEE Robot. Autom. Mag., vol. 17, no. 2, pp. 44–54, Jun. 2010.
[5] S. M. Khansari-Zadeh and A. Billard, “Learning stable nonlinear dynamical systems with Gaussian mixture models,” IEEE Trans. Robot., vol. 27, no. 5, pp. 943–957, Oct. 2011.
[6] S. M. Khansari-Zadeh and A. Billard, “Learning control Lyapunov function to ensure stability of dynamical system-based robot reaching motions,” Robot. Auton. Syst., vol. 62, no. 6, pp. 752–765, 2014.
[7] S. M. Khansari-Zadeh and O. Khatib, “Learning potential functions from human demonstrations with encapsulated dynamic and compliant behaviors,” Auton. Robots, vol. 41, no. 1, pp. 45–69, 2017.
[8] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, “Dynamical movement primitives: learning attractor models for motor behaviors,” Neural Comput., vol. 25, no. 2, pp. 328–373, 2013.
[9] V. Sindhwani, S. Tu, and M. Khansari, “Learning contracting vector fields for stable imitation learning,” 2018. [Online]. Available: arXiv:1804.04878.
[10] D.-X. Zhou, “Derivative reproducing properties for kernel methods in learning theory,” J. Comput. Appl. Math., vol. 220, nos. 1–2, pp. 456–463, 2008.
[11] C. Carmeli, E. De Vito, and A. Toigo, “Vector–valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem,” Anal. Appl., vol. 4, no. 4, pp. 377–408, 2006.
[12] A. A. Ahmadi and B. El Khadir, “Learning dynamical systems with side information,” Mach. Learn. Res., vol. 1, p. 10, Jun. 2020.
[13] S. M. Richards, F. Berkenkamp, and A. Krause, “The Lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems,” in Proc. Conf. Robot Learn., vol. 87, 2018, pp. 466–476.
[14] F. Berkenkamp, R. Moriconi, A. F. Schoellig, and A. Krause, “Safe learning of regions of attraction for uncertain, nonlinear systems with Gaussian processes,” in Proc. Conf. Decis. Control, 2016, pp. 4661–4666.
[15] W. Wang, P. Yu, L. Lin, and T. Tong, “Robust estimation of derivatives using locally weighted least absolute deviation regression,” J. Mach. Learn. Res., vol. 20, no. 60, pp. 1–49, 2019.
[16] J. Puyouquet, Convex Optimization in Normed Spaces: Theory, Methods and Examples. Cham, Switzerland: Springer, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-13710-0.
[17] M. Chaves, T. Eissinger, and F. Allgower, “Bistable biological systems: A characterization through local compact input-to-state stability,” IEEE Trans. Autom. Control, vol. 53, pp. 87–100, Jan. 2008.
[18] I. R. Manchester and J.-J. E. Slotine, “Control contraction metrics: Convex and intrinsic criteria for nonlinear feedback design,” IEEE Trans. Autom. Control, vol. 62, no. 6, pp. 3046–3053, Jun. 2017.