Convex Nonparametric Formulation for Identification of Gradient Flows

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

Abstract—We develop a nonparametric identification method for nonlinear gradient-flow dynamics. In these systems, the vector field is the gradient field of a potential energy function. This fundamental fact about the dynamics is a structural prior knowledge and a constraint in the proposed identification method. While the nature of the identification problem is an estimation in the space of functions, we derive an equivalent finite dimensional formulation, which is a convex optimization in the form of a quadratic program. This provides scalability and the opportunity for utilizing recently developed large-scale optimization solvers. The central idea is representing the energy function as a difference of two convex functions and learning these convex functions jointly. Based on necessary and sufficient conditions for function convexity, the identification problem is formulated, and then, the existence, uniqueness and smoothness of the solution is addressed. We also illustrate and evaluate the method numerically with two demonstrative examples.

Index Terms—Nonlinear system identification, gradient flows, convex optimization, quadratic program.

I. INTRODUCTION

ONLINEAR dynamics are ubiquitous in nature and widely used for modeling various phenomena in science and engineering [1], [2]. These models are either derived from first principles or by fitting and estimation methods. The latter employ techniques in optimization, learning theory and system identification for deriving the model from the available measurement data. Meanwhile, in many cases modeling goes beyond fitting a nonlinear dynamics to the data. We may additionally need to incorporate specific properties into the model, which are inherent in the nature of system. These properties includes stability, passivity, positivity, or other possible characteristics of the system. For linear dynamics, many of these properties are already considered [3]–[6].

For the nonlinear systems, a similar line of research has received extensive attention in the past decade [7]–[9], e.g., identifying stabilizable non-autonomous dynamics [7], [8], and learning the dynamics subject to sparsity of the dynamic modes of system [9]. For the purpose of imitation learning, the dynamics modeled in [10] are based on Gaussian mixtures and hidden Markov models. A similar approach is employed in [11] with a global stability guarantee. In [12], a convex quadratic potential energy and a linear dissipative field are considered with respect to each data point, and the dynamics are modeled as a functional weighted sum of the gradient forces and the dissipative fields. In [13], the dynamics are modeled as weakly nonlinear differential equations with a linear part for capturing the baseline behavior together with more complicated coupling dynamics for considering more complex phenomena.

An interesting class of nonlinear dynamics are gradient-flows defined as the negative of gradient of a potential energy function with classical examples of electric and gravitational forces and Langevin dynamics [14]. This type of dynamics appears in modeling physical systems like robots. Therefore, an identification method considering this feature of dynamics can be profitable in the relevant cases such as in learning inverse dynamics [15]. The structural properties of gradient flows can be used as a prior knowledge, as well as a constraint in the identification problem. In [16], a method is introduced, based on incremental stability and contraction analysis, which is generally computationally demanding and also not suitable for learning gradient flows of non-convex energy functions. Moreover, their arguments lack concise theoretical guarantees.

We introduce a nonparametric identification method for the gradient flow dynamics. The problem is originally formulated as a minimization of a fitting metric over the hypothesis space of convex functions. Following this, we derive an equivalent finite dimensional convex optimization problem. For the sake of more transparent discussion and ease of notation, first, the case of convex energy functions are discussed and then, the case of general energy functions is addressed. Finally, the method is numerically illustrated and evaluated.

II. NOTATIONS AND PRELIMINARIES

The set of natural numbers, the set of non-negative integers, the set of real numbers, n-dimensional Euclidean space and the space of n by m real matrices are denoted by \( \mathbb{N} \), \( \mathbb{Z}_{\geq} \), \( \mathbb{R} \), \( \mathbb{R}^n \), and \( \mathbb{R}^{n \times m} \) respectively. The identity matrix and zero vector in the Euclidean space are denoted by \( \mathbb{I} \) and \( \mathbf{0} \) respectively. The set of positive definite matrices in \( \mathbb{R}^{n \times n} \) is denoted by \( \mathbb{S}_{++}^n \). For matrices \( X, Y \in \mathbb{R}^{n \times n} \), we write \( X \succeq Y \) if \( X - Y \in \mathbb{S}_{++}^n \). For a
set $A \subseteq \mathbb{R}^n$, the convex hull of $A$ is denoted by $\text{conv}(A)$. The Euclidean norm and the inner product on $\mathbb{R}^n$ are respectively denoted by $\| \cdot \|$ and $\langle \cdot, \cdot \rangle$. For a function $f$, $\nabla f$ and $\nabla^2 f$ are the gradient and Hessian of $f$ respectively. For a convex function $\varphi : U \subseteq \mathbb{R}^n \to \mathbb{R}$, the subgradient of $f$ at $x$ is denoted by $\partial f(x)$ and defined as the set of vectors $\xi \in \mathbb{R}^n$ satisfying the inequality $\varphi(y) - \varphi(x) \geq \langle \xi, y - x \rangle$, for all $y \in U$. Let $Y$ be a set and $C$ be a subset of $Y$. The function $I_C$ is defined as $I_C(y) = 0$, if $y \in C$ and $I_C(y) = \infty$, otherwise.

III. PROBLEM STATEMENT

Let $U$ be a simply-connected open subset of $\mathbb{R}^n$ and $\varphi : U \to \mathbb{R}$ be an unknown function. We call $\varphi$ the potential energy function or simply, energy function. A conservative vector field corresponding to $\varphi$ is induced over the space, and the corresponding dynamics are defined as

$$\dot{x} = f(x) := -\nabla \varphi(x).$$

(1)

Starting from $x_0 \in U$ at time $t = 0$, the vector field generates a trajectory which is denoted by $x(\cdot; x_0)$. Consider points $x_0, \ldots, x_{nt}$ and corresponding trajectories. For any $i = 1, \ldots, nt$, let trajectory $x(\cdot; x_i)$ be sampled at time instants $0 \leq t_1^i < \cdots < t_{ni}^i$, where $n_i \in \mathbb{N}$. Let $x_k^i$ denote measured $x(t_k^i; x_i)$, for $1 \leq k \leq n_i$. The measurements are given to a pre-processing unit for estimating the time derivative of $x(t_i^i; x_i)$ at the sampling time instants. This unit might work using various estimation tools, e.g., a finite-difference procedure, or a non-linear regression method together with an analytical calculation of derivatives. For derivative estimation examples see [17] and the references therein. Let these estimations be denoted by $y_k^i$, for $1 \leq k \leq n_i$. Note that $y_k^i$ is approximately equal to $f(x_k^i)$. Accordingly, we can introduce a set of data, denoted by $D$, which contains data pairs $(x_k^i, y_k^i)$. Indeed, $D$ is defined as $\{(x_j, y_j) \mid j \in I_t\}$, where $n_t := \sum_{1 \leq i \leq nt} n_i, I_t := \{1, \ldots, n_t\}$ and, for simplicity of notation, the superscripts are dropped. Note that once $D$ is provided, we are not considering any particular time-ordering for the data points.

Problem: Given the set of data $D$, estimate the unknown vector field $f$ in (1).

Remark 1. This problem is a nonlinear system identification where structural prior knowledge is provided in form of (1). This can be generalized to the case of differential inclusions.

IV. CONVEX ENERGY FUNCTIONS

In this section, we consider the case of convex energy functions. In the following, we relax the differentiability assumption of energy function for the sake of generality.

Assumption 1. The potential energy function $\varphi$ is convex.

Let $\Phi$ denote the set of convex functions defined over $\mathbb{R}^n$. Considering the data $D$, we define the loss function for the estimation problem, denoted by $L_{\Phi, D}$, as the sum of squared errors. In other words, $L_{\Phi, D} : \Phi \times \mathbb{R}^{n \times nt} \to \mathbb{R}$ is a function such that for any given convex function $\varphi \in \Phi$ and vectors $\xi_1, \ldots, \xi_{nt}$, the value of $L_{\Phi, D}(\varphi, \xi)$ is defined as

$$L_{\Phi, D}(\varphi, \xi) := \sum_{i=1}^{nt} \|y_i + \xi_i\|^2 + \sum_{i=1}^{nt} I_{\partial \varphi(x_i)}(\xi_i).$$

(2)

where $\xi \in \mathbb{R}^{n \times nt}$ is $\xi := [\xi_1^T \ldots \xi_{nt}^T]$ and $y_i$ is the estimated value for $f(x_i)$, for any $i \in I_t$. Note that for any $\varphi \in \Phi$, we have that $\partial \varphi(x) \neq \emptyset$, for any $x$, and also, $\partial \varphi(x) = \{\nabla \varphi(x)\}$, when $\varphi$ is differentiable at $x$ [18]. Accordingly, the estimation problem is naturally defined as

$$\min_{\xi_1, \ldots, \xi_{nt} \in \mathbb{R}^n, \varphi \in \Phi} \sum_{i=1}^{nt} \|y_i + \xi_i\|^2,$$

s.t. $\xi_i \in \partial \varphi(x_i), \quad i = 1, \ldots, nt$. (3)

Note that optimization problem (3) is over the set $\Phi$, a cone in the space of functions which is an infinite-dimensional space. We provide a tractable approach for solving (3).

A. Towards Finite-Dimensional Formulation

For any convex function $\varphi$, the following holds [18]

$$\varphi(y) - \varphi(x) \geq \langle \xi, y - x \rangle, \quad \forall x, y \in \mathbb{R}^n, \forall \xi \in \partial \varphi(x).$$

(4)

Motivated by this property of convex functions, we introduce the following optimization problem

$$\min_{\xi \in \mathbb{R}^{n \times nt}, \theta \in \mathbb{R}^n} \sum_{i=1}^{nt} \|y_i + \xi_i\|^2,$$

s.t. $\theta_j - \theta_i \geq \langle \xi_i, x_j - x_i \rangle, \quad i, j \in I_t$. (5)

Define the vector $\theta \in \mathbb{R}^n$ as $\theta := [\theta_1 \ldots \theta_{nt}]$. Let $K$ be the feasible set in (5). Considering the optimization (5), one can define a loss function $L_{K, D} : \mathbb{R}^n \times \mathbb{R}^{n \times nt} \to \mathbb{R}$ as

$$L_{K, D}(\theta, \xi) := \sum_{i=1}^{nt} \|y_i + \xi_i\|^2 + I_K(\theta, \xi).$$

(6)

The next theorem presents the connection between (3) and (5).

Theorem 1 (Equivalency Theorem). i) Let $(\varphi, \xi)$ be a solution of (3). Then, (5) has a solution $(\theta, \xi)$ such that $L_{K, D}(\theta, \xi) = \Phi_{\varphi}(\varphi, \xi)$. ii) Conversely, if $(\theta, \xi)$ is a solution of (5), then (3) has a solution $(\varphi, \xi)$ such that $L_{K, D}(\varphi, \xi) = L_{K, D}(\theta, \xi)$ and $\xi \in \partial \varphi(x_i)$, for all $i \in I_t$. Proof i): For any $i \in \{1, \ldots, nt\}$, define $\theta_i = \varphi(x_i)$. From (4), one can easily see that $(\theta, \xi) \in K$, and subsequently, $I_K(\theta, \xi) = 0$. Therefore, we have

$$L_{K, D}(\varphi, \xi) = \sum_{i=1}^{nt} \|y_i + \xi_i\|^2 = L_{K, D}(\varphi, \xi).$$

(7)

Let (5) have a feasible point $(\bar{\theta}, \bar{\xi})$ such that $L_{K, D}(\bar{\theta}, \bar{\xi}) < L_{K, D}(\theta, \xi)$. Therefore, we have

$$\sum_{i=1}^{nt} \|y_i + \bar{\xi}_i\|^2 = L_{K, D}(\bar{\theta}, \bar{\xi}) < L_{K, D}(\theta, \xi) = \sum_{i=1}^{nt} \|y_i + \xi_i\|^2.$$

(8)

Let function $\bar{\varphi} : \mathbb{R}^n \to \mathbb{R}$ be defined as

$$\bar{\varphi}(x) := \max_{1 \leq i \leq n_t} \langle \bar{\xi}_i, x - x_i \rangle + \bar{\theta}_i, \quad \forall x \in \mathbb{R}^n.$$

(9)

One can easily see that $\bar{\varphi}$ is a convex function, i.e., $\bar{\varphi} \in \Phi$. Define set-valued map $\bar{I} : \mathbb{R}^n \rightrightarrows \{1, \ldots, n_t\}$ as

$$\bar{I}(x) = \{i \in \{1, \ldots, n_t\} \mid \bar{\varphi}(x) = \langle \bar{\xi}_i, x - x_i \rangle + \bar{\theta}_i\}.$$ (10)

For any $x \in \mathbb{R}^n$, we know that [18]

$$\partial \bar{\varphi}(x) = \text{conv} \{\bar{\xi}_i \mid i \in \bar{I}(x)\}.$$ (11)

Since $(\bar{\theta}, \bar{\xi}) \in K$, for any $i = 1, \ldots, nt$, we have

$$\bar{\theta}_i \geq \langle \bar{\xi}_i, x_i - x_j \rangle + \bar{\theta}_j, \quad \forall j = 1, \ldots, n_t.$$ (12)
Therefore, from (9) and (12), one can see that
\[ \tilde{\psi}(x_i) \geq \langle \tilde{\xi}_i, x_i - x_i \rangle + \tilde{\theta}_i = \hat{\theta}_i \geq \max_{i=1,...,n} \langle \tilde{\xi}_i, x_i - x_i \rangle + \hat{\theta}_i = \tilde{\psi}(x_i). \]  
(13)

Subsequently, due to (11), \( \tilde{\xi}_i \in \partial \tilde{\psi}(x_i) \) and \( \tilde{\psi}(x_i) \) is a feasible point for (3). From (2), (6), (7) and (8), we have
\[ L_{\Phi, \mathcal{D}}(\tilde{\psi}, \tilde{\xi}) \leq L_{\Phi, \mathcal{D}}(\tilde{\theta}, \tilde{\xi}) \leq L_{\Phi, \mathcal{D}}(\theta, \xi), \]
which is a contradiction and \( L_{\Phi, \mathcal{D}}(\tilde{\theta}, \tilde{\xi}) \leq L_{\Phi, \mathcal{D}}(\theta, \xi) \). Therefore, \( (\theta, \xi) \) is a solution of (5) and the proof is concluded. 

Part ii): Let \((\theta, \xi)\) be a solution of (5). Define \( \psi : \mathbb{R}^n \rightarrow \mathbb{R} \) as \( \psi(x) := \max_{i=1,...,n} \langle \xi_i, x - x_i \rangle + \theta_i \). Note that \( \psi \) is a convex function, i.e., \( \psi \in \Phi \). Define set-valued map \( \mathcal{I} : \mathbb{R}^n \rightarrow \{1, ..., n\} \) similar to (10). For any \( x \in \mathbb{R}^n \), we have \( \partial \psi(x) = \text{conv}(\{\xi_i | i \in \mathcal{I}(x)\}) \). Since \((\theta, \xi) \in \mathcal{K} \), based on a similar argument to the proof of part (ii), we have that \( \tilde{\xi}_i \in \partial \psi(x_i) \), for any \( i = 1, ..., n \). Subsequently, we have
\[ L_{\Phi, \mathcal{D}}(\psi, \xi) = \sum_{i=1}^{n} \|y_i + \xi_i\|^2 = L_{\Phi, \mathcal{D}}(\theta, \xi). \]  
(14)

Now, let (3) have a feasible point \((\tilde{\psi}, \tilde{\xi})\) such that
\[ L_{\Phi, \mathcal{D}}(\tilde{\psi}, \tilde{\xi}) < L_{\Phi, \mathcal{D}}(\psi, \xi). \]
For any \( i = 1, ..., n \), define \( \tilde{\theta}_i = \tilde{\psi}(x_i) \). Since \( \tilde{\psi} \) is a convex function, due to (4), one can see that \( (\tilde{\theta}, \tilde{\xi}) \in \mathcal{K} \). From (2), (6) and (14), we have
\[ L_{\Phi, \mathcal{D}}(\theta, \xi) = L_{\Phi, \mathcal{D}}(\tilde{\theta}, \tilde{\xi}) < L_{\Phi, \mathcal{D}}(\psi, \xi) = L_{\Phi, \mathcal{D}}(\theta, \xi), \]
which is a contradiction. This shows that \((\theta, \xi)\) is a solution of (3). (3) concludes the proof of part (ii).

**Theorem 2 (Main Theorem).** Optimization problem (3) admits a solution in form of
\[ \psi(x) := \max_{i=1,...,n} \langle \xi_i, x - x_i \rangle + \theta_i, \]  
(15)
where \((\theta, \xi)\) is a solution of (5). Moreover, we have \( L_{\Phi, \mathcal{D}}(\psi, \xi) \leq L_{\Phi, \mathcal{D}}(\theta, \xi) \) and \( \tilde{\xi}_i \in \partial (\psi)(x_i) \), for all \( i = 1 \).

**Proof.** Optimization problem (5) is a quadratic program. Since \( \mathcal{K} \) is a non-empty polyhedral cone, (5) has a solution, denoted by \((\theta, \xi)\). Therefore, due to Theorem 1, optimization problem (3) admits a solution in form of (15). The rest of the theorem follows from Theorem 1 and the given proof.

Based on Theorem 2, one can solve (5) instead of the equivalent set.\( \Theta^* \) is the unique solution, denoted by \((\theta^*, \xi^*)\). Moreover, \( \lim_{\lambda \rightarrow 0}(\theta^*, \xi^*) \) exists and is equal to \((\theta^*, \xi^*)\) where \( \theta^* := \text{argmin}_{\theta \in \Theta^*} \|\theta\|^2 \). Also, \( \lim_{\lambda \rightarrow \infty}(\theta, \xi) \) exists and equals to \((0, \xi^*)\) where \( \xi^* \) is the unique solution of \( \min_{(\theta, \xi) \in \mathcal{K}} \sum_{i=1}^{n} \|y_i + \xi_i\|^2 + \lambda \|\theta\|^2 \). 

**Proof.** One can easily see that \((0, 0) \in \mathcal{K} \) and \( \mathcal{V}^2 \mathcal{J}_\lambda \geq \lambda \mathcal{I} \). Therefore, (19) is an optimization problem with a strongly convex cost function and non-empty closed and convex feasible set. Therefore, (19) has a unique solution. Similarly, since \( \Theta^* \) is non-empty, closed and convex, \( \theta^* := \text{argmin}_{\theta \in \Theta^*} \|\theta\|^2 \) is well-defined and exists uniquely. From the definition of \((\theta^*, \xi^*) \) and \((\theta, \xi)\), one can easily see that for any \( \lambda > 0 \), we have \( \sum_{i=1}^{n} \|y_i + \xi_i\|^2 + \lambda \|\theta_i\|^2 \leq \sum_{i=1}^{n} \|y_i + \xi_i\|^2 + \lambda \|\theta\|^2 \), and subsequently, it holds that \( \|\theta\| \leq \|\theta^*\| \). Similarly, since \((0, 0) \in \mathcal{K} \), one can see that \( \|\xi^*\|^2 + \|\xi^*\|^2 \leq 4 \sum_{i=1}^{n} \|y_i\|^2 \). Now, define set \( \mathcal{C} \subset \mathbb{R}^n \times \mathbb{R}^n \) as \( \mathcal{C} := \{(\theta, \xi) | \|\theta\| \leq \|\theta^*\|, \|\xi\| \leq 2(\sum_{i=1}^{n} \|y_i\|^2)^{\frac{1}{2}}\} \), which is a compact and convex set. Define sets \( \mathcal{Z}_0 := \text{argmin}_{(\theta, \xi) \in \mathcal{C}} \mathcal{J}(\theta, \xi) \) and \( \mathcal{Z} := \text{argmin}_{(\theta, \xi) \in \mathcal{C}} \mathcal{J}(\theta, \xi) \). We know that \( \mathcal{K} \cap \mathcal{C} \) is a compact set and \( \mathcal{J}_\lambda(\theta, \xi) \) is a continuous function with respect to \((\theta, \xi)\). Therefore, due to Maximum Theorem [19], the set-valued map \( \lambda \mapsto \mathcal{Z}_\lambda \) is upper hemi-compact with non-empty and compact values. Moreover, one has \( \mathcal{Z}_0 = \{(\theta^*, \xi^*)\} \) and \( \mathcal{Z} = \{(\theta^*, \xi^*)\} \). Subsequently, from the upper hemi-compactness of the map \( \lambda \mapsto \mathcal{Z}_\lambda \), we have \( \lim_{\lambda \rightarrow 0}(\theta, \xi) \) exists and equals to \((\theta^*, \xi^*)\). Replacing \( \lambda \) with \( \lambda^{-1} \), one can show the last part of the theorem by repeating same steps of the proof.

Given \( \lambda > 0 \), we can define our estimator as following
\[ \psi_\lambda(x) := \max_{i=1,...,n} \langle \xi_{\lambda,i}, x - x_i \rangle + \theta_{\lambda,i}, \]  
(20)
where \((\theta_{\lambda}, \xi_{\lambda})\) is the unique solution of (19).
C. Smoothing the Estimator

Given $\lambda > 0$, the smooth version of (20) is defined as
\[
\psi_{\lambda, \tau}(x) = \tau \log \left( \frac{1}{n} \sum_{1 \leq i \leq n} \exp \left( \frac{1}{\tau} - (\xi_{\lambda,i} - x) + \theta_{\lambda,i} \right) \right),
\]
(21)

Theorem 5 [18]. Let the log-sum-exp function $\ell : \mathbb{R}^n \to \mathbb{R}$ be defined as $\ell(x) := \log(\sum_{i=1}^{n} e^{x_i})$, for any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. This function is an analytical and convex function. The gradient and Hessian of $\ell$ are
\[
\nabla \ell(x) = \frac{z}{\|z\|^2}, \quad \text{and} \quad \nabla^2 \ell(x) = \frac{zz^T}{\|z\|^4} - \frac{zz^T}{\|z\|^2},
\]
where $z = [e^{x_1}, \ldots, e^{x_n}]^T$. Moreover, the following holds
\[
\max_{1 \leq i \leq n} x_i \leq \ell(x) \leq \max_{1 \leq i \leq n} x_i + \ln n, \quad \forall x \in \mathbb{R}^n.
\]

This function is used to define a smooth approximation to $\varphi_\lambda$ in (20). More precisely, let $\mathbb{X}_\lambda$ be the matrix defined as $\mathbb{X}_\lambda := [\xi_{\lambda,1}, \xi_{\lambda,2}, \ldots, \xi_{\lambda,n}] \in \mathbb{R}^{n \times n}$, and, for $i = 1, \ldots, n$, define $\eta_{\lambda,i}$ as $\eta_{\lambda,i} := \theta_i - (\xi_{\lambda,i} - x_i)$ and subsequently, let $\eta_\lambda$ be the vector defined as $\eta_\lambda := [\eta_{\lambda,1}, \ldots, \eta_{\lambda,n}]^T$. Subsequently, one can see that
\[
\psi_{\lambda, \tau}(x) = \tau \ell \left( \frac{1}{\tau} \left( \mathbb{X}_\lambda^T x + \eta_\lambda \right) \right) - \ln n_\lambda,
\]
(23)

The next corollary motivates the use of $\psi_{\lambda, \tau}$ as a smooth approximant to $\varphi_\lambda$.

Corollary 1. For any $\tau > 0$, the function $\psi_{\lambda, \tau}$, defined in (21), is convex and analytical. Moreover, we have
\[
\nabla \psi_{\lambda, \tau}(x) = \mathbb{X}_\lambda \nabla \ell(x), \quad \text{and} \quad \nabla^2 \psi_{\lambda, \tau}(x) = \mathbb{X}_\lambda \nabla^2 \ell(x)
\]
where $z = [\exp (\{[\xi_{\lambda,n}, x - x_j] + \theta_{\lambda,j}\})]_{j=1}^{n}$. Also, we have
\[
\varphi_\lambda(x) - \tau \ln n_\lambda \leq \psi_{\lambda, \tau}(x) \leq \varphi_\lambda(x), \quad \forall x \in \mathbb{R}^n.
\]
(24)

Corollary 2. By taking $\tau$ small enough, $\psi_{\lambda, \tau}$ as a uniform approximant of $\varphi_\lambda$. More precisely, let $\epsilon$ be an arbitrary positive real scalar and let $\tau < \frac{\epsilon}{\|z\|}$. Then (24) shows that $|\varphi_\lambda(x) - \psi_{\lambda, \tau}(x)| < \epsilon$, for any $x \in \mathbb{R}^n$.

D. Further Extensions

The proposed estimation strategy can be extended to other similar settings by suitably modifying the main estimation problem (3) and the corresponding finite-dimensional version (5). For example, if we know that the energy function $\varphi$ belongs to the class of $\mu$-strongly convex functions, denoted by $\Phi_\mu$, then the estimation problem (3) is adapted to
\[
\min_{\xi \in \mathbb{R}^{n \times n}, \theta \in \mathbb{R}^n} \sum_{i=1}^{n} \|y_i + \xi_i\|^2, \quad \text{s.t.} \quad \xi_i \in \partial \varphi(x_i), \quad i, j \in I_k.
\]
(25)
and, the optimization problem (5) is modified to
\[
\min_{\xi \in \mathbb{R}^{n \times n}, \theta \in \mathbb{R}^n} \sum_{i=1}^{n} \|y_i + \xi_i\|^2, \quad \text{s.t.} \quad \theta_j - \theta_i \geq \langle \xi_i, x_j - x_i \rangle + \frac{\mu}{2} \|x_j - x_i\|^2, \quad i, j \in I_k.
\]

Another case is when an equilibrium of (1) is known which is denoted by $x_0$. In order to incorporate this knowledge, the estimation problem (3) should be modified to the following,
\[
\min_{\xi \in \mathbb{R}^{n \times n}, \theta \in \mathbb{R}^n} \sum_{i=1}^{n} \|y_i + \xi_i\|^2, \quad \text{s.t.} \quad \theta_j - \theta_i \geq \langle \xi_i, x_j - x_i \rangle + \frac{\mu}{2} \|x_j - x_i\|^2, \quad i, j \in I_k.
\]
(26)

Without loss of generality, we can assume that $\varphi(x_0) = 0$. Accordingly, one can set $\theta_0 = 0$ and $\xi_0 = 0$. Therefore, we modify the optimization problem (5) as following
\[
\min_{\xi \in \mathbb{R}^{n \times n}, \theta \in \mathbb{R}^n} \sum_{i=1}^{n} \|y_i + \xi_i\|^2, \quad \text{s.t.} \quad \theta_j - \theta_i \geq \langle \xi_i, x_j - x_i \rangle, \quad i, j \in I_k \cup \{0\}.
\]
(27)

Finally, when we know that $\varphi$ is a concave function, the adaptation is straightforward due to convexity of $-\varphi$. For all of these cases, the regularization and the smoothing procedures follow the same lines as before.

V. GENERAL ENERGY FUNCTIONS

In this section, we consider general energy functions. For the sake of generality, the differentiability assumption of the energy function is relaxed initially. The next theorem plays a key role in the formulation of the estimation problem.

Theorem 6 [20]. (i) Let $\Omega \subseteq \mathbb{R}^n$ be a convex set and $\varphi : \Omega \to \mathbb{R}$ be a $C^2(\Omega, \mathbb{R})$ function with bounded Hessian, i.e., $\sup_{x \in \Omega} \|\nabla^2 \varphi(x)\| < \infty$. Then, there exist convex functions $\varphi^1, \varphi^2 : \Omega \to \mathbb{R}$ such that $\varphi(x) = \varphi^1(x) - \varphi^2(x)$, for any $x \in \Omega$. (ii) Moreover, if $\Omega$ is convex and compact, then the Hessian is bounded and $\varphi$ has a convex decomposition.

The loss function for the estimation problem, $L_{\varphi, \varphi} : \Phi \times \Phi \times \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} \to \mathbb{R}$ is defined as
\[
L_{\varphi, \varphi}(\varphi^1, \varphi^2, \xi^1, \xi^2) := \sum_{i=1}^{n} \left\| y_i + \xi^1_i - \xi^2_i \right\|^2
\]
\[
+ \sum_{i=1}^{n} T_{\partial \varphi^1(\xi^1)}(\xi^1_i) + \sum_{i=1}^{n} T_{\partial \varphi^2(\xi^2)}(\xi^2_i),
\]
(28)
for any pair of convex functions $\varphi^1, \varphi^2 \in \Phi$ and vectors $\xi^1_1, \ldots, \xi^1_n, \xi^2_1, \ldots, \xi^2_n \in \mathbb{R}^n$, $\xi^1_1, \xi^2_1 \in \mathbb{R}^{n \times n}$ where $\xi^1_1, \xi^2_1 \in \mathbb{R}^{n \times n}$ are column vectors respectively defined as $\xi^1 := [\xi^1_1 \ldots \xi^1_n]^T$ and $\xi^2 := [\xi^2_1 \ldots \xi^2_n]^T$. Accordingly, the estimation problem is defined as
\[
\min_{\xi^1 \in \mathbb{R}^{n \times n}, \xi^2 \in \mathbb{R}^{n \times n}} \sum_{i=1}^{n} \left\| y_i + \xi^1_i - \xi^2_i \right\|^2, \quad \text{s.t.} \quad \xi^1_i \in \partial \varphi^1(x_i), \quad i, j \in I_k, \quad k = 1, 2.
\]
(29)

Analogous to the previous section, we can introduce a finite-dimensional formulation as following
\[
\min_{(\theta^1, \xi^1) \in \mathbb{R}^{n \times n}, k=1,2} \sum_{i=1}^{n} \left\| y_i + \xi^1_i - \xi^2_i \right\|^2, \quad \text{s.t.} \quad \theta^1_j - \theta^1_i \geq \langle \xi^1_i, x_j - x_i \rangle, \quad i, j \in I_k, \quad k = 1, 2.
\]
(30)
where $\theta^{(1)}, \theta^{(2)} \in \mathbb{R}^n$ are defined respectively as $\theta^{(1)} = [\theta_1^{(1)}, \ldots, \theta_n^{(1)}]^T$ and $\theta^{(2)} = [\theta_1^{(2)}, \ldots, \theta_n^{(2)}]^T$.

Considering this optimization problem, we define a loss function $\mathcal{L}_{\mathcal{C}, \mathcal{D}} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as $\mathcal{L}_{\mathcal{C}, \mathcal{D}}(\theta^{(1)}, \theta^{(2)}, \xi^{(1)}, \xi^{(2)}) = \sum_{i=1}^n ||y_i + \xi_i^{(1)} - \xi_i^{(2)}||^2 + \mathcal{I}_\mathcal{C}(\theta^{(1)}, \xi^{(1)}) + \mathcal{I}_\mathcal{C}(\theta^{(2)}, \xi^{(2)})$. With lines of proof similar to those in Section IV, we formalize the connection between optimization problems (29) and (30).

**Theorem 7.** Optimization problem (29) admits a solution as

$$
\phi(x) \doteq \phi^{(1)}(x) - \phi^{(2)}(x) = \max_{1 \leq s \leq n} (\xi^{(1)}_s, x - x_h) + \theta^{(1)}_s - \max_{1 \leq s \leq n} (\xi^{(2)}_s, x - x_h) + \theta^{(2)}_s,
$$

where $(\theta^{(1)}, \theta^{(2)}, \xi^{(1)}, \xi^{(2)})$ is a solution of (30). Moreover, $\mathcal{L}_{\mathcal{C}, \mathcal{D}}(\phi^{(1)}, \phi^{(2)}, \xi^{(1)}, \xi^{(2)}) = \mathcal{L}_{\mathcal{C}, \mathcal{D}}(\theta^{(1)}, \theta^{(2)}, \xi^{(1)}, \xi^{(2)})$, and $\xi^{(1)}_i \in \partial \phi^{(1)}(x_i)$, $\xi^{(2)}_i \in \partial \phi^{(2)}(x_i)$, for all $i \in I_s$.

As in Section IV, regularization can be used for imposing uniqueness in the estimation and improving numerical stability. Using the same arguments as those given in the proof of Theorem 3, one can obtain a similar conclusion and subsequently show that the difference $\xi^{(1)}_i - \xi^{(2)}_i$ is unique and the potential non-uniqueness of the solution is due to the other terms. Consequently, one can introduce the regularized cost function $\mathcal{J}_\mathcal{C} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as $\mathcal{J}_\mathcal{C}(\theta^{(1)}, \theta^{(2)}, \xi^{(1)}, \xi^{(2)}) = \sum_{i=1}^n ||y_i + \xi^{(1)}_i - \xi^{(2)}_i||^2 + \lambda(||\theta^{(1)}||^2 + ||\theta^{(2)}||^2 + ||\xi^{(1)}||^2 + ||\xi^{(2)}||^2)$, and solve the following regularized optimization problem

$$
\min_{(\theta^{(1)}, \xi^{(1)}) \in K, (\theta^{(2)}, \xi^{(2)}) \in K} \mathcal{J}_\mathcal{C}(\theta^{(1)}, \theta^{(2)}, \xi^{(1)}, \xi^{(2)}),
$$

where is the regularization weight. Similar to Theorem 4, one can show that for any $\lambda > 0$, the regularized optimization problem (31) has a unique solution, denoted by $(\theta^{(1)}_\lambda, \theta^{(2)}_\lambda, \xi^{(1)}_\lambda, \xi^{(2)}_\lambda)$. Consequently, we define $\phi^{(1)}_\lambda$ and $\phi^{(2)}_\lambda$ similar to (20), and thus, the estimation is defined as $\phi^{(1)}_\lambda(x) \doteq \phi^{(1)}_\lambda(x) - \phi^{(2)}_\lambda(x)$ where $\phi^{(1)}_\lambda(x) \doteq \max_{1 \leq s \leq n} (\xi^{(1)}_s, x - x_h) + \theta^{(1)}_\lambda$, for $j = 1, 2$. While the current form of regularization is for the sake of minimum norm solution, one may include other forms for promoting various interesting features. For example, one may penalize the non-smoothness of the potential function by introducing a regularization function as the norm of total variation of the potential function or its derivatives. Note that the uniqueness of the corresponding solution is guaranteed if the objective function, i.e., the fitting error plus the regularization function, is a strongly convex function. Similar to the previous section, we can smooth this estimator using the log-sum-exp function. In this regard, let $\phi^{(1)}_\lambda$ and $\phi^{(2)}_\lambda$ be defined as in (21), and then define the smooth estimator, denoted by $\phi_{\lambda, \tau}$, as

$$
\phi_{\lambda, \tau}(x) = \phi^{(1)}_{\lambda, \tau}(x) - \phi^{(2)}_{\lambda, \tau}(x).
$$

Note that Corollary 1 and Corollary 2 are valid for $\phi^{(1)}_{\lambda, \tau}$ and $\phi^{(2)}_{\lambda, \tau}$. Moreover, we have the following corollary for $\phi_{\lambda, \tau}$.

**Corollary 3.** Let $\epsilon > 0$ and $\tau < \frac{1}{2\lambda}$. Then, due to (24), we have $|\phi_\lambda(x) - \phi_{\lambda, \tau}(x)| < \epsilon$, for any $x \in \mathbb{R}^n$. In other words, we can uniformly approximate $\phi_\lambda$ with $\phi_{\lambda, \tau}$ to an arbitrary accuracy by taking $\tau$ sufficiently small.

**Remark 2.** Let $\{h_k\}_{k=1}^p$ be a set of given vector fields. Then, the proposed method can be extended to the case where the dynamics is in the form of $f(x) = -\nabla \phi(x) + \sum_{k=1}^p a_k h_k(x)$. To estimate $f$, we modify optimization problem (29) to

$$
\min_{(\theta^{(1)}, \theta^{(2)} \in \mathbb{R}^n, \sigma \in \Phi_{\alpha_1, \ldots, \alpha_p}} \sum_{i=1}^n ||y_i + \bar{\xi}^{(1)}_i - \bar{\xi}^{(2)}_i - \sum_{k=1}^p a_k h_k(x)||^2,
$$

s.t. $\bar{\xi}^{(j)}_i \in \partial \phi^{(j)}(x_i)$, $i \in I_s$, $j = 1, 2$, and then, apply previously discussed adaptations.

**VI. NUMERICAL EXPERIMENTS**

In this section, we discuss numerical examples for the illustration and comparisons of the presented method.

**Example 1.** Consider the non-convex energy function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ defined as $\varphi(x_1, x_2) := ax_1^2 + bx_1x_2 + ax_2^2 - cx_1^4 - cx_2^4$, where $a = 0.7, b = -0.5$ and $c = 0.15$. The graph of energy function is shown in Fig. 1. We consider 18 initial points in $\Omega := [-1.9, 1.9] \times [-1.9, 1.9] \subset \mathbb{R}^2$. Following this, we take 118 measurement samples with additive white Gaussian noise of zero mean and standard deviation $\sigma_w = 0.01$. Fig. 1 shows the initial points and the sampled points by red bullets and black bullets, respectively. From these points, 80% are randomly chosen as the data set for estimating $f(x) = -\nabla \varphi(x)$. By solving (31), a close to minimum norm estimation is obtained (see Theorem 4 and Section V). Then, using the log-sum-exp function a smooth version of $\varphi$ is derived as in (32). The parameters $\lambda$ and $\tau$ are tuned via a standard cross-validation procedure using the remaining 20% of the data points, i.e., the parameters are chosen such that the generalization error on this 20% of the data is minimized. The results are $\lambda = 10^{-8}$ and $\tau = 0.16$. Following this, we estimate $\partial_{x_1} \varphi(x_1, x_2)$ and $\partial_{x_2} \varphi(x_1, x_2)$ due to the gradient of the smoothed function and Corollary 1. Given the ground truth and the data set $\mathcal{D}$, the coefficient of determination, also known as $R$-squared, is 92.4%.

**Example 2.** In this example, we numerically evaluate the accuracy of the proposed method and provide comparisons with [16] and [11]. To this end, we consider potential energy function $\varphi(x) = \frac{1}{2} \ln(a x^2 + 1)$ where $a = 10$. Initializing the dynamics at $t = 0$ to $x_0 = \pm 1.5$, we obtain the corresponding responses of system and then, sample them with sampling time $\Delta t = 0.05$ for time interval $[0, 7]$. In order to investigate and compare the sensitivity of the estimation methods,
we consider various levels of uncertainties in the set of data \( \mathcal{D} \) and evaluate the resulting errors of estimation. This numerical analysis is illustrated in Fig. 2 (right). Also, for SNR \( \approx 40 \) dB, the results are shown in Fig. 2 (left). One can see that the proposed method shows better performance.

Discussion: Methods CVF-CF and CVF-GS [16] are based on incremental stability and contraction analysis. The estimation problem is formulated as a convex optimization problem with \( n \) fully coupled \( n \times n \) linear matrix inequalities (LMI) forcing the Jacobian of vector field to be a Hurwitz matrix at the sampling points. Accordingly, the optimization problem becomes computationally demanding, specially for identification of gradient flows where informative sets of data are essentially large. Moreover, \( f = -\nabla \phi \) together with the LMI constraints leads to the convexity of \( \phi \) near the sampling locations. This feature limits the application of these methods, particularly when the potential energy is non-convex. This issue is clear from the monotone form of the estimated functions in Fig. 2. In [11] the dynamics is modeled as a Gaussian mixture where the parameters of model are fitted by minimizing the mean square error which is a non-convex optimization. Meanwhile, the method presented here is based on a quadratic program which is computationally tractable. The methods in [11] and [16] demand exhaustive hyperparameter tuning, while the estimation problem (30) is completely non-parametric and (31) has one hyperparameter.

VII. Conclusion

We have introduced a nonparametric system identification method for nonlinear systems with gradient-flow dynamics. This fact is a structural prior knowledge which is used as a constraint in the proposed method. Initially, the identification problem is formulated as a minimization of the fitting error over the hypothesis space of convex functions. Then, a tractable problem equivalent formulation is derived as a finite dimensional quadratic program. This formulation is based on two central ideas: representing the energy function as a difference of two convex functions, and a necessary and sufficient condition for convexity. The existence, uniqueness and smoothness of the solution is addressed. Finally, we have illustrated and compared the method via numerical examples.

One interesting future direction is providing theoretical analysis for the impact of errors and uncertainties on the estimation. Furthermore, for enforcing continuity and smoothness, one may reformulate the problem where the hypothesis space of potential functions is a subset of a Sobolev space. Considering other regularizations for penalizing the discontinuities and non-smoothness are also interesting directions.

REFERENCES

[1] S. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Cambridge, MA, USA: Perseus Books, 2001.
[2] T. Q. Chen, Y. Robozova, J. Bettencourt, and D. K. Dvavena, “Neural ordinary differential equations,” in *Proc. 32nd Conf. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6571–6583.
[3] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, “Kernel methods in system identification, machine learning and function estimation: A survey,” *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.
[4] J. Umenberger, J. Wägberg, I. R. Manchester, and T. B. Schön, “Maximum likelihood identification of stable linear dynamical systems,” *Automatica*, vol. 96, pp. 280–292, Oct. 2018.
[5] I. Goethals, T. Van Gestel, J. Suykens, P. Van Dooren, and B. De Moor, “Identification of positive real models in subspace identification by using regularization,” *IEEE Trans. Autom. Control*, vol. 48, no. 10, pp. 1843–1847, Oct. 2003.
[6] M. Khosravi and R. S. Smith, “Kernel-based identification of positive systems,” in *Proc. IEEE 58th Conf. Decis. Control*, Nice, France, 2019, pp. 1740–1745.
[7] Y. Sattar and S. Oymak, “Non-asymptotic and accurate learning of non-linear dynamical systems,” 2020. [Online]. Available: arXiv:2002.08538.
[8] 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.
[9] E. Kaiser, J. N. Kutz, and S. L. Brunton, “Sparse identification of nonlinear dynamics for model predictive control in the low-data limit,” *Proc. Roy. Soc. A*, vol. 474, no. 2219, 2018, Art. no. 20180335.
[10] S. Calinon, F. D’hallewin, 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.
[11] S. M. Khansari-Zadeh and A. Billard, “Learning stable nonlinear dynamical systems with Gaussian mixture models,” *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 943–957, Oct. 2011.
[12] 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.
[13] A. J. Ipspeert, 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, Feb. 2013.
[14] K. Bakshi, P. Grover, and E. A. Theodorou, “On mean field games for agents with Langevin dynamics,” *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 4, pp. 1451–1460, Dec. 2019.
[15] D. Nguyen-Tuong, J. Peters, M. Seeger, and B. Schölkopf, “Learning inverse dynamics: A comparison,” in *Proc. Eur. Symp. Artif. Neural Netw.*, 2008, pp. 13–18.
[16] V. Sindhwani, S. Tu, and M. Khansari, “Learning contracting vector fields for stable imitation learning,” 2018. [Online]. Available: M:arXiv:1804.04878.
[17] 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.
[18] Y. Nesterov, *Lectures on Convex Optimization*. Cham, Switzerland: Springer, 2018. [Online]. Available: https://doi.org/10.1007/978-3-19-91578-4
[19] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker’s Guide*. Berlin, Germany: Springer, 2006. [Online]. Available: https://doi.org/10.1007/978-3-540-29587-9
[20] A. L. Yuille and A. Rangarajan, “The concave-convex procedure,” *Neural Comput.*, vol. 15, no. 4, pp. 915–936, 2003.