Identification of nonlinear systems and optimality analysis in Sobolev spaces

Carlo Novara [*], **Angelo Nicolí** [*], **Giuseppe C. Calafiore** [*]

[*] *Dip. di Elettronica e Telecomunicazioni, Politecnico di Torino*
Corso Duca degli Abruzzi, 24 - 10129 Torino, Italy.
Email: {carlo.novara,angelo.nicoli,giuseppe.calafiore}@polito.it

Abstract: In this paper, we propose a novel approach for the identification from data of an unknown nonlinear function together with its derivatives. This approach can be useful, for instance, in the context of nonlinear system identification for obtaining models that are more reliable than the traditional ones, based on plain function approximation. Indeed, models identified by accounting for the derivatives can provide a better performance in several tasks, such as multi-step prediction, simulation, and control design. We also develop an optimality analysis, showing that models derived using this approach enjoy suitable optimality properties in Sobolev spaces. We finally demonstrate the effectiveness of the approach with a numerical example.

Keywords: Model identification, nonlinear systems, optimality, identification for control

1. INTRODUCTION

Consider a nonlinear discrete-time system, represented in the following input-output regression form:

$$y_{t+1} = f_o(x_t) + \xi_{t+1} \tag{1}$$
$$x_t = (y_t, \dots, y_{t-m_u+1}, u_t, \dots, u_{t-m_u+1})$$

where $u_t \in U \subset \mathbb{R}^{n_u}$ is the input, $y_t \in \mathbb{R}^{n_y}$ is the output, $\xi_t \in \varXi \subset \mathbb{R}^{n_\xi}$ is a disturbance and $t \in \mathbb{Z}$ is the discrete time index. The sets U and $\varXi$ are compact with non-empty interior. The *regression function* f_o is supposed to be unknown. In this paper, we consider the problem of obtaining from a batch of experimental data an estimate $\hat{f}$ of f_o such that (*i*) $\hat{f}$ approximates f_o, and (*ii*) the first derivatives of $\hat{f}$ approximate the first derivatives of f_o.

The key motivation for considering this problem is the following. In general, when estimating a regression model that is to be used for control, e.g., of the type $\hat{y}_{t+1} = \hat{f}(y_t, \dots, y_{t-m_u+1}, u_t, \dots, u_{t-m_u+1})$, it is of paramount importance to capture the sensitivities of the output with respect to the commands $u_t, \dots, u_{t-m_u+1}$, and these are given, to first order approximation, by the derivatives of $\hat{f}$ w.r.t. these variables. Failing to get these sensitivities with sufficient precision may result in a model that responds to commands in a poor way.

The literature appears to be quite scarce on the topic of approximating from data a function and its derivatives. The existing methods are based on different classes of approximators, including radial basis functions (Mai-Duy and Tran-Cong, 2003), neural networks (Xie and Cao, 2011; Pukrittayakamee et al., 2011; Avrutskiy, 2018), and deep neural networks (Czarnecki et al., 2017). The numerical results presented in these papers clearly show that using the information about the function derivatives leads to significant improvements of the model accuracy and generalization capabilities. This literature is interesting and effective in showing the potential of techniques relying on derivative identification. However, only a limited number of works carry out a theoretical analysis about the approximation properties of these techniques (Hornik et al., 1990; Xie and Cao, 2011; Czarnecki et al., 2017), and the provided results are often non-constructive, in the sense that they just prove existence of the required approximating function. Also, we observe that the existing techniques allow for the identification of a model, but they do not provide a description of the uncertainty associated with this model and its predictions.

In this paper, we propose a novel identification approach addressing all the mentioned issues. The approach allows the identification of a function together with its derivatives, and it is completely based on convex optimization. We develop a theoretical optimality analysis, showing that models obtained using the proposed approach enjoy certain optimality properties in Sobolev spaces. We finally present a numerical example, concerned with multi-step prediction of the Chua chaotic circuit. This example shows that the approach may provide significantly more accurate and reliable models than the traditional ones based on plain function approximation (i.e., identified without considering the derivatives).

2. NOTATION AND PRELIMINARIES

A column vector $x \in \mathbb{R}^{n_x \times 1}$ is denoted by $x = (x_1, \dots, x_{n_x})$. A row vector $x \in \mathbb{R}^{1 \times n_x}$ is denoted by $x = [x_1, \dots, x_{n_x}] = (x_1, \dots, x_{n_x})^\top$, where $\top$ indicates the transpose. The ℓ_p norm of a vector $x = (x_1, \dots, x_{n_x})$ is defined as usual and denoted with $\|x\|_p$. The 2-norm (maximum singular value norm) of a matrix $\Phi \in \mathbb{R}^{m \times n}$ is denoted by $\|\Phi\|_2$, and the ∞-norm is denoted by $\|\Phi\|_\infty \doteq \max_{i=1,\dots,m} \sum_{j=1}^{n} |\Phi_{ij}|$. The $\mathcal{L}_p$ norm of a function with domain $X \subseteq \mathbb{R}^{n_x}$ and codomain in $\mathbb{R}$, is defined as $\|f\|_p \doteq \left[\int_X \|f(x)\|_p^p dx\right]^{\frac{1}{p}}$, for $p \in (1, \infty)$, and as

$\operatorname{ess\,sup}_{x\in X} \|f(x)\|_\infty$ for $p = \infty$. These norms give rise to the well-known ℓ_p and $\mathcal{L}_p \equiv \mathcal{L}_p(X)$ Banach spaces. The $\mathcal{S}_{1p}$ Sobolev norm of a differentiable function with domain $X \subseteq \mathbb{R}^{n_x}$ and codomain in $\mathbb{R}$, is defined as $\|f\|_{\mathcal{S}p} \doteq \sum_{i=0}^{n_x} \|f^{(i)}\|_p$, where $f^{(i)} \doteq f$ for $i = 0$, and $f^{(i)} \doteq \frac{\partial f}{\partial x_i}$ for $i > 0$. Note that the superscript (i), with $i > 0$, here denotes the partial derivative of a function with respect to the i-th variable, and not the i-th order derivative. The Sobolev norm gives rise to the $\mathcal{S}_{1p} \equiv \mathcal{S}_{1p}(X)$ Sobolev space, also denoted in the literature with W_{1p} or $W_{1,p}$.

Definition 1. The Sobolev space $\mathcal{S}_{1p}(X)$ is the set of all functions $f \in \mathcal{L}_p(X)$ such that, for every $i > 0$, the derivative $f^{(i)}$ exists and $f^{(i)} \in \mathcal{L}_p(X)$: $\mathcal{S}_{1p}(X) \doteq \{f : f^{(i)} \in \mathcal{L}_p(X),\, i = 0, \ldots, n_x\}$. □

Sobolev norms (and related spaces) involving higher order derivatives can also be found in the literature. The concept of weak derivative, which is a generalization of the standard derivative, is often used. In this paper, the interest is for the case of first order standard derivatives.

3. PROBLEM FORMULATION

Consider a function $f_o \in \mathcal{S}_{1p}(X)$, taking values $z = f_o(x)$, where $x \in X \subset \mathbb{R}^{n_x}$, X is a compact set, and $z \in \mathbb{R}$. Suppose that f_o is not known, but a set of noise-corrupted input-output data from the unknown function is available:

$$D = \left\{\tilde{x}_k, \left\{\tilde{z}_k^i\right\}_{i=0}^{n_x}\right\}_{k=1}^{L} \tag{2}$$

where $\tilde{x}_k \in X$ are the measurements of the function argument, $\tilde{z}_k^0 \equiv \tilde{z}_k$ are the measurements of the function output and $\tilde{z}_k^i$, $i > 0$, are the measurements of the i-th partial derivative output. The data (2) can be described by

$$\tilde{z}_k^i = f_o^{(i)}(\tilde{x}_k) + d_k^i,\ i = 0, \ldots, n_x,\ k = 1, \ldots, L, \tag{3}$$

where d_k^i are noises and $d_k^0 \equiv d_k$. If the data are generated by the system (1), we have that $\tilde{z}_k^0 \equiv \tilde{z}_k = \tilde{y}_{k+1}$, and the noise terms account for the disturbance ξ_t and possible measurement errors.

We remark that in real-world applications, only the output of the function is usually measured, while the outputs of the derivatives may not be available. This situation is dealt with in Novara et al. (2019), where an algorithm is presented for estimating the derivative output samples $\tilde{z}_k^i$, $i > 0$, from the input-output function samples $\tilde{x}_k$ and $\tilde{z}_k$.

Now, assume that the noise sequences $d^i = (d_1^i, \ldots, d_L^i)$ are unknown but bounded:

$$\|d^i\|_q \leq \mu^i \tag{4}$$

where $\|\cdot\|_q$ is a vector ℓ_q norm and $0 \leq \mu^i < \infty$. In the case $q = 2$, it can be convenient to write μ^i as $\mu^i = \sqrt{L}\breve{\mu}^i$, with $0 \leq \breve{\mu}^i < \infty$. In some situations, the noise bounds μ^i are known from the physical knowledge about the system of interest and the involved sensors. In other situations, these bounds are not known and have to be estimated from the available data. An algorithm is provided in Novara et al. (2019) for performing this estimation.

In this paper, we consider the problem of identifying from the data (2) an "accurate" approximation $\hat{f}$ of the unknown function f_o, such that also the derivatives $\hat{f}^{(i)}$, $i > 0$, of $\hat{f}$ are "accurate" approximations of the derivatives $f_o^{(i)}$, $i > 0$, of f_o. The accuracy is measured by means of the following identification error $e(\hat{f}) \doteq \|f_o - \hat{f}\|_{\mathcal{S}p}$, where $\|\cdot\|_{\mathcal{S}p}$ is a Sobolev norm. In other words, we are looking for an approximation of the unknown function f_o in the $\mathcal{S}_{1p}$ Sobolev space. Besides the goal of obtaining such an approximation, we also aim at evaluating guaranteed estimate bounds for f_o.

A parametrized structure is adopted for the approximating function:

$$\hat{f}(x) = \sum_{j=1}^{N} a_j \phi_j(x) \tag{5}$$

where $\phi_j \in \mathcal{S}_{1p}(X)$ are given basis functions and $a_j \in \mathbb{R}$ are coefficients to be identified. The choice of the basis functions is clearly an important step of the identification process, see, e.g., (Sjöberg et al., 1995; Novara et al., 2011). In several cases, the basis functions are known from the physical knowledge of the system of interest. In other cases the basis functions are known a priori to belong to some "large" set of functions, see, e.g., the examples presented in Section 6 and in (Novara, 2011). In yet other cases, the basis functions are not known a priori and their choice can be carried out by considering the numerous options available in the literature (e.g., Gaussian, sigmoidal, wavelet, polynomial, trigonometric, etc.); see (Sjöberg et al., 1995) for a discussion on the main features of the most used basis functions and guidelines for their choice.

Problem 1. From the data set D in (2), identify an estimate $\hat{f}$ of the form (5), such that:

(i) the Sobolev identification error $e(\hat{f})$ is small;
(ii) the estimate is equipped with guaranteed uncertainty bounds on the unknown function f_o and its derivatives.

The concept of "small" identification error will be formalized in Section 5, according to suitable optimality criteria.

In the reminder of the paper, for numerical conditioning reasons, we assume that the components of x in $z = f_o(x)$ have similar ranges of variation. This assumption can always be met through a suitable rescaling of the components.

4. IDENTIFICATION METHODS

In this section we propose two methods for solving Problem 1, both based on convex optimization. In Section 5 it will be shown that functions identified by means of these methods enjoy suitable optimality properties. In this section, we suppose that the derivative output samples $\tilde{z}_k^i$, $i > 0$ are available or estimated using the algorithm in Novara et al. (2019).

A simple yet fundamental observation is that the approximating function (5) and its derivatives are given by

$$\hat{f}^{(i)}(x) = \sum_{j=1}^{N} a_j \phi_j^{(i)}(x),\ i = 0, \ldots, n_x. \tag{6}$$

On the basis of this observation we can present the first identification method.

Method 1.

(1) Define

$$\tilde{z}^i \doteq \begin{bmatrix} \tilde{z}_1^i \\ \vdots \\ \tilde{z}_L^i \end{bmatrix}, \; \Phi^i \doteq \begin{bmatrix} \phi_1^{(i)}(\tilde{x}_1) & \cdots & \phi_N^{(i)}(\tilde{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1^{(i)}(\tilde{x}_L) & \cdots & \phi_N^{(i)}(\tilde{x}_L) \end{bmatrix}. \quad (7)$$

(2) Estimate the vector $a = (a_1, \ldots, a_N)$ of model coefficients in (6) by solving the following convex optimization problem:

$$a = \arg \min_{\alpha \in \mathbb{R}^N} \|\alpha\|_r \quad (8)$$

$$\text{s.t. } \|\tilde{z}^i - \Phi^i \alpha\|_q \leq \mu^i, \; i = 0, \ldots, n_x, \quad (9)$$

where the integers r, q indicate suitable vector norms.

The rationale behind this method can be explained as follows: the constraints (9) ensure that the resulting model (6) is consistent with the available information on the noises corrupting the data. If the optimization problem is not feasible, it means that either the chosen basis function set is not sufficiently rich or the noise bounds $\|d^i\|_q \leq \mu^i$ are too small. The minimization of the coefficient vector ℓ_r norm in (8) is carried out for regularization reasons, allowing also to limit the issue of overfitting. Typical norms that can be used are the ℓ_2 and ℓ_1 norms. In particular, the ℓ_1 norm allows one to obtain a sparse coefficient vector a (see, e.g., (Fuchs, 2005; Tibshirani, 1996; Tropp, 2006; Donoho et al., 2006)), resulting in a low-complexity model. This is an important property, especially in view of the model implementation on real-time processors.

We now present the second identification method.

Method 2.

(1) Define $\tilde{z}^i$ and Φ^i as in (7).

(2) Estimate the vector $a = (a_1, \ldots, a_N)$ of model coefficients in (6) by solving the following convex optimization problem:

$$a = \arg \min_{\alpha \in \mathbb{R}^N} \sum_{i=0}^{n_x} \lambda^i \|\tilde{z}^i - \Phi^i \alpha\|_q^2 + \Lambda \|\alpha\|_r \quad (10)$$

where the integers r, q indicate suitable vector norms, and $\lambda^i \geq 0, \Lambda \geq 0$ are given weights.

Problem (10) is aimed at minimizing a tradeoff between the model fitting error on the identification data and a regularization term. For $r = 1$, (10) is a Lasso problem, see, e.g., (Tibshirani, 1996); for $r = 2$, it becomes a classical Ridge regression problem, see, e.g., (Gruber, 1998). Note that, for suitable values of the parameters μ^i, λ^i and Λ, the optimization problems (8) and (10) are equivalent to each other.

Remark 1. It is worth to stress the fact that Method 1 and Method 2 are here considered in terms of the guarantees they provide for the ensuing models, and that this paper's contribution lies in the specific models that lead to Sobolev space identification through Method 1 and Method 2, and in their analysis, and *not* in the actual numerical solution of problems in (8) or (10). These problems indeed have a well-known regularized regression structure, and a pletora of efficient numerical methods already exist for their solution. ⋆

5. OPTIMALITY ANALYSIS

In Section 4, two identification methods have been presented, allowing us to derive parameterized approximations of the unknown function f_o. In this section, following a Set Membership approach (Milanese and Vicino, 1991), (Milanese et al., 1996), (Schweppe, 1973), (Chen and Gu, 2000), (Milanese and Novara, 2011), (Sznaier et al., 2009), we show that such approximations enjoy suitable optimality properties in Sobolev spaces. The analysis and results developed here are extensions to Sobolev spaces of those regarding approximation in $\mathcal{L}_p$ spaces presented in (Milanese and Novara, 2004, 2011).

Consider that the function f_o and its derivatives are unknown, while instead we have the experimental information given by (2) and (3), and the prior information given by the inclusion $f_o \in \mathcal{S}_{1p}(X)$ and the noise bounds $\|d^i\|_q \leq \mu^i$. It follows that $f_o \in \text{FFS}_\mathcal{S}$, where $\text{FFS}_\mathcal{S}$ is the so-called Feasible Function Set, defined below.

Definition 2. The Feasible Function Set $\text{FFS}_\mathcal{S}$ is defined as $\text{FFS}_\mathcal{S} \doteq \{f \in \mathcal{S}_{1p}(X) : ||\tilde{z}^i - f^{(i)}(\tilde{x})||_q \leq \mu^i, \, i = 0, \ldots, n_x\}$, where $f^{(i)}(\tilde{x}) \doteq (f^{(i)}(\tilde{x}_1), \ldots, f^{(i)}(\tilde{x}_L))$. □

In words, the Feasible Function Set is the set of all functions consistent with the prior assumptions and with the available data. The Feasible Function Set thus summarizes all the experimental and a-priori information that can be used for identification. If at least a function exists that is consistent with the assumptions and the data (i.e., if $\text{FFS}_\mathcal{S} \neq \emptyset$), we say that the assumptions are validated. Otherwise (i.e., if $\text{FFS}_\mathcal{S} = \emptyset$), we say that the assumptions are falsified; see (Milanese et al., 1996; Chen and Gu, 2000).

Definition 3. The prior assumptions are considered validated if $\text{FFS}_\mathcal{S} \neq \emptyset$. □

The following theorem gives a sufficient condition for prior assumption validation.

Theorem 1. $\text{FFS}_\mathcal{S} \neq \emptyset$ if the optimization problem (8)-(9) is feasible.

Proof. See Novara et al. (2019). □

If the optimization problem (8)-(9) is not feasible, it means that either the chosen basis function set is not sufficiently rich or the noise bounds $\|d^i\|_q \leq \mu^i$ are too small. In the case where reliable noise bounds are available, a sufficiently rich basis function set has to be found, considering the numerous options available in the literature (e.g., Gaussian, sigmoidal, wavelet, polynomial, trigonometric). If no basis functions are found for which the optimization problem is feasible, a relaxation of the noise bounds is needed.

In the reminder of the paper, it is assumed that the prior assumptions are true and, consequently, $f_o \in \text{FFS}_\mathcal{S}$. Under this assumption, for a given approximation $\hat{g}$ of f_o, a tight bound on the identification error $e(\hat{g})$ is given by the following worst-case error.

Definition 4. We define the worst-case identification error as $\text{WE}(\hat{g}, \text{FFS}_\mathcal{S}) \doteq \sup_{f \in \text{FFS}_\mathcal{S}} \|f - \hat{g}\|_{\mathcal{S}p}$, where $\| \cdot \|_{\mathcal{S}p}$ is the Sobolev norm. □

An optimal approximation is defined as a function f_{op} which minimizes the worst-case approximation error.

Definition 5. An approximation f_{op} is $\text{FFS}_\mathcal{S}$-optimal if $\text{WE}(f_{op}, \text{FFS}_\mathcal{S}) = \inf_{\hat{g}} \text{WE}(\hat{g}, \text{FFS}_\mathcal{S}) \doteq \mathcal{R}(\text{FFS}_\mathcal{S})$, where $\mathcal{R}(\text{FFS}_\mathcal{S})$ is called the *radius of information* and is the minimum worst-case error that can be achieved on the basis of the available prior and experimental information. □

In other words, an optimal approximation is the best approximation that can be found on the basis of the available prior and experimental information (this information is summarized by the Feasible Function Set). Finding optimal approximations is in general hard and sub-optimal solutions can be looked for. In particular, approximations called almost-optimal are often considered in the literature, see, e.g., (Traub et al., 1988), (Milanese et al., 1996).

Definition 6. An approximation f_{ao} is $\mathrm{FFS}_{\mathcal{S}}$-almost-optimal if $\mathrm{WE}(f_{\mathrm{ao}}, \mathrm{FFS}_{\mathcal{S}}) \leq 2 \inf_{\hat{g}} \mathrm{WE}(\hat{g}, \mathrm{FFS}_{\mathcal{S}}) = 2\mathcal{R}(\mathrm{FFS}_{\mathcal{S}})$. □

The following result gives sufficient conditions under which an approximation (possibly obtained by the methods of Section 4) is almost-optimal.

Theorem 2. Assume that:
i) the optimization problem (8)-(9) is feasible.
ii) the approximation $\hat{f}$ given in (5)-(6) has coefficients a_j satisfying inequalities (9).
Then, the approximation $\hat{f}$ is $FFS_{\mathcal{S}}$-almost-optimal.

Proof. See Novara et al. (2019). □

This theorem shows that an approximation obtained by Method 1 is always almost-optimal. Instead, an approximation obtained by Method 2 is almost-optimal if its coefficients satisfy inequalities (9).

In Novara et al. (2019), the optimality analysis is extended to the case where an additional Lipschitz continuity assumption is made. This assumption allows us to prove stronger optimality properties with respect to those discussed above. Moreover, in Novara et al. (2019), tight uncertainty bounds are derived, for the unknown function f_o and its derivatives $f_o^{(i)}$, $i = 1, \ldots, n_x$. These bounds quantify the modeling error and the prediction uncertainty. They can be useful in real-world applications for several purposes, such as robust control design (Freeman and Kokotovic, 1996), (Qu, 1998), prediction interval evaluation (Milanese and Novara, 2005), and fault detection (Novara, 2016). Based on the uncertainty bounds, an algorithm for the estimation of the noise bounds μ^i is also presented in Novara et al. (2019).

6. EXAMPLE: MULTI-STEP PREDICTION FOR THE CHUA CHAOTIC CIRCUIT

The Chua circuit is a simple electronic circuit showing a chaotic behavior, see (Chua et al., 1986). It is composed of two capacitors, an inductor, a locally active resistor and a nonlinear resistor. The circuit continuous-time state equations are the following:

$$\begin{aligned} \dot{x}_1 &= \alpha(x_2 - x_1 - \rho(x_1)) \\ \dot{x}_2 &= x_1 - x_2 + x_3 + u + \xi^c \\ \dot{x}_3 &= -\beta x_2 - R x_3 \\ y &= x_1 \end{aligned} \tag{11}$$

where the states $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$ represent the voltages across the capacitors, $x_3 \in \mathbb{R}$ the current through the inductor, $u \in \mathbb{R}$ is an external input, $y \in \mathbb{R}$ is the system output, $\xi^c \in \mathbb{R}$ is a disturbance, and $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ and $R \in \mathbb{R}$ are parameters. In this example, the following nonlinear resistor characteristic and parameter values are assumed: $\rho(x_1) = -1.16x_1 + 0.041x_1^3$, $R = 0.1$, $\alpha = 10.4$, $\beta = 16.5$. With this parameter values and nonlinearity, the system exhibits a chaotic behavior and thus prediction is an extremely hard task.

The system (11), discretized via the forward Euler method, can be written in the following input-output regression form:

$$\begin{aligned} y_t &= b_1 y_{t-1} + b_2 y_{t-2} + b_3 y_{t-3} \\ &\quad + b_4 \rho(y_{t-1}) + b_5 \rho(y_{t-2}) + b_6 \rho(y_{t-3}) \\ &\quad + b_7 u_{t-2} + b_8 u_{t-3} + \xi_t \end{aligned} \tag{12}$$

where ξ_t is a noise accounting for the disturbance ξ^c in (11) and b_i are suitable parameters. Equivalently, it can be written in the form (1), with $x_t = (y_t, y_{t-1}, y_{t-2}, u_{t-1}, u_{t-2})$.

The system (11) has been implemented in Simulink. The input u was simulated as a normally distributed random signal with zero mean and standard deviation (std) 1. The disturbance ξ^c was simulated as a normally distributed random signal with zero mean. Two std values were considered for this disturbance: 0.01 and 0.05. For each of these std values, two simulations of duration 60 s were carried out and, correspondingly, two set of data of the form (2) were collected with a sampling time $T_s = 0.01$ s, corresponding to an experiment length $L = 6000$ for every dataset. The first dataset was used for model identification, the second one for model validation.

For each std value of the disturbance ξ^c, the following prediction models were identified from the identification dataset.

- *One-step predictor identified not using any derivative information (P1_NOD).* The predictor P1_NOD is given by

$$\begin{aligned} y_{t+1} &= \hat{f}\,(x_t) \\ x_t &= (y_t, y_{t-1}, y_{t-2}, u_{t-1}, u_{t-2}) \end{aligned} \tag{13}$$

  where $\hat{f}$ is of the form (5). A basis function set composed of multivariate monomials has been used, defined as

$$\{\phi_j\}_{j=1}^N = \{\prod_{i=1}^{n_x} x_{i,t}^{\alpha_i - 1}; \alpha_i = 1, 2; i = 1, \ldots, n_x\} \tag{14}$$

  where $x_{i,t}$ is the ith component of x_t and $n_x = 5$. This set consists of $N = 2^{n_x} = 32$ basis functions. The coefficients a_j in (5) were identified by Method 2, with $q = 2$, $r = 1$, $\lambda^0 = 1$, $\lambda^i = 0$, $i > 0$, and $\Lambda = 50$.

- *One-step predictor identified using the true derivative values (P1_D).* The predictor P1_D is of the form (13). The basis functions are the same as those used in (13). The true derivative values computed from (12) were used to construct the vector $\tilde{z}^i$, $i > 0$, in (7). The coefficients a_j in (5) were identified by Method 2, with $q = 2$, $r = 1$, $\lambda^0 = 1$, $\lambda^i = 200$, $i > 0$, and $\Lambda = 50$.

- *One-step predictor identified using the estimated derivative values (P1_ED).* The predictor P1_ED is of the form (13). The basis functions are the same as those used in (13). The derivative values estimated by the algorithm in Novara et al. (2019) were used to construct the vector $\tilde{z}^i$, $i > 0$, in (7). The coefficients

a_j in (5) were identified by Method 2, with $q = 2$, $r = 1$, $\lambda^0 = 1$, $\lambda^i = 200$, $i > 0$, and $\Lambda = 50$.

- *Direct multi-step predictor identified not using any derivative information (PK_NOD).* The predictor PK_NOD is given by

$$\begin{aligned} y_{t+k} &= \hat{f}\left(x_t\right) \\ x_t &= \left(y_t, y_{t-1}, y_{t-2}, u_{t+k-2}, u_{t+k-3}, \ldots, u_{t-2}\right) \end{aligned} \tag{15}$$

  where $\hat{f}$ is of the form (5) and $k \in \{3,5,7\}$. The basis function set is defined as in (14), with $n_x = 4 + k$. This set consists of $N = 2^{4+k}$ basis functions. The coefficients a_j in (5) were identified by Method 2, with $q = 2$, $r = 1$, $\lambda^0 = 1$, $\lambda^i = 0$, $i > 0$, and $\Lambda = 50$.

- *Direct multi-step predictor identified using the estimated derivative values (PK_ED).* The predictor PK_ED is of the form (15). The basis functions are the same as those used in (15). The derivative values estimated by the algorithm in Novara et al. (2019) were used to construct the vector $\tilde{z}^i$, $i > 0$, in (7). The coefficients a_j in (5) were identified by Method 2, with $q = 2$, $r = 1$, $\lambda^0 = 1$, $\lambda^i = 200$, $i > 0$, and $\Lambda = 50$.

For each std value of the disturbance ξ^c (std $\in \{0.01, 0.05\}$), the identified models were tested on the validation set in the task of k-step ahead prediction, with $k \in \{3,5,7\}$. The k-step prediction of models P1_NOD, P1_D and P1_ED was computed by iterating k times equation (13). The k-step prediction of models PK_NOD and PK_ED was computed directly using equation (15).

The results of these tests are summarized in Tables 1 and 2, where the Root Mean Square prediction Errors RMSE$_k$ are reported, for $k \in \{3,5,7\}$ and std $\in \{0.01, 0.05\}$. Figure 1 shows the true system output and the 3-step prediction of the model PK_ED (in the case where std = 0.05) for a portion of the validation set. The uncertainty bounds computed according to the approach given in Novara et al. (2019) are also reported in the figure. Note that these results were obtained using Method 2. Similar results can be obtained using Method 1 (they are not reported here for the sake of brevity).

The main observation arising from these results is that the models identified by the proposed method, using the information about the derivatives, are significantly more accurate (about one order of magnitude) than those identified not using this information. A second observation is that the models identified using the estimated derivative values show a performance similar to those identified using the true derivative values. A third observation (important in general but less important than the other two in the context considered in this paper) is that the direct k-step predictors are in general more accurate than the iterated 1-step predictors.

REFERENCES

Avrutskiy, V. (2018). Enhancing approximation abilities of neural networks by training derivatives. *arXiv:1712.04473v2.*

Chen, J. and Gu, G. (2000). *Control-Oriented System Identification: An H_∞ Approach.* John Wiley & Sons, New York.

Predictors	RMSE$_3$	RMSE$_5$	RMSE$_7$
P1_NOD	6.21e-02	1.03e-01	1.45e-01
P1_D	5.55e-03	1.27e-02	2.29e-02
P1_ED	3.60e-03	1.01e-02	2.09e-02
PK_NOD	6.01e-02	9.97e-02	1.41e-01
PK_ED	5.69e-04	8.73e-04	1.87e-03

Table 1. Validation set; std = 0.01; $k \in \{3,5,7\}$. RMSE prediction errors.

Predictors	RMSE$_3$	RMSE$_5$	RMSE$_7$
P1_NOD	6.18e-02	1.03e-01	1.44e-01
P1_D	5.59e-03	1.29e-02	2.33e-02
P1_ED	3.56e-03	1.01e-02	2.09e-02
PK_NOD	5.96e-02	9.90e-02	1.39e-01
PK_ED	6.22e-04	1.07e-03	2.17e-03

Table 2. Validation set; std = 0.05; $k \in \{3,5,7\}$. RMSE prediction errors.

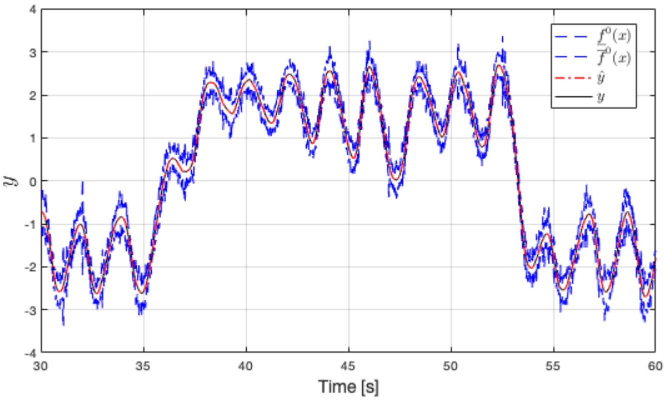

Fig. 1. Validation set (a portion); std = 0.05. 3-step prediction of model PK_ED and related uncertainty bounds.

Chua, L., Komuro, M., and Matsumoto, T. (1986). The double scroll family. *IEEE Transactions on Circuits and Systems*, 33(11), 1072–1118.

Czarnecki, W., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. (2017). Sobolev training for neural networks. *arXiv:1706.04859v3.*

Donoho, D., Elad, M., and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1), 6 – 18. doi:10.1109/TIT.2005.860430.

Findeisen, R., Allgower, F., and Biegel, L. (2007). Assessment and future directions of nonlinear model predictive control. In *Lecture Notes in Control and Information Sciences*. Springer.

Freeman, A. and Kokotovic, V. (1996). *Robust Nonlinear Control Design.* Birkhuser, Boston.

Fuchs, J. (2005). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10), 3601 –3608. doi: 10.1109/TIT.2005.855614.

Goodwin, G., Yuz, J., Aguero, J., and Cea, M. (2010). Sampling and sampled-data models. In *American Control Conference, plenary lecture.* Baltimore, MD, USA.

Gruber, M. (1998). *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators.*

CRC Press.

Grune, L. and Pannek, J. (2011). Nonlinear model predictive control - theory and algorithms. In *Communications and Control Engineering*. Springer.

Hornik, K., Stinchcombe, M., and White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*. doi:10.1016/0893-6080(90)90005-6.

Ljung, L. (1999). *System identification: theory for the user*. Prentice Hall, Upper Saddle River, N.J.

Magni, L., Raimondo, D., and Allgower, F. (2009). Nonlinear model predictive control - towards new challenging applications. In *Lecture Notes in Control and Information Sciences*. Springer.

Mai-Duy, N. and Tran-Cong, T. (2003). Approximation of function and its derivatives using radial basis function networks. *Applied Mathematical Modelling*. doi: 10.1016/S0307-904X(02)00101-4.

Manzano, J., Limon, D., de la Peñ, D.M., and Calliess, J. (2018). Robust data-based model predictive control for nonlinear constrained systems. *IFAC-PapersOnLine*, 51(20), 505 – 510.

Milanese, M., Norton, J., Lahanier, H.P., and Walter, E. (1996). *Bounding Approaches to System Identification*. Plenum Press.

Milanese, M. and Novara, C. (2004). Set membership identification of nonlinear systems. *Automatica*, 40/6, 957–975. doi:10.1016/j.automatica.2004.02.002.

Milanese, M. and Novara, C. (2005). Set membership prediction of nonlinear time series. *IEEE Transactions on Automatic Control*, 50(11), 1655–1669. doi: 10.1109/TAC.2005.858693.

Milanese, M. and Novara, C. (2011). Unified set membership theory for identification, prediction and filtering of nonlinear systems. *Automatica*, 47(10), 2141–2151. doi: 10.1016/j.automatica.2011.03.013.

Milanese, M. and Vicino, A. (1991). Optimal algorithms estimation theory for dynamic systems with set membership uncertainty: an overview. *Automatica*, 27, 997–1009.

Novara, C. (2011). Sparse identification of nonlinear functions and parametric set membership optimality analysis. In *American Control Conference*. San Francisco, California.

Novara, C. (2016). Sparse set membership identification of nonlinear functions and application to fault detection. *International Journal of Adaptive Control and Signal Processing*, 30(2), 206–223.

Novara, C., Formentin, S., Savaresi, S., and Milanese, M. (2016). Data-driven design of two degree-of-freedom nonlinear controllers: the D2-IBC approach. *Automatica*, 72, 19–27.

Novara, C. and Milanese, M. (2019). Control of mimo nonlinear systems: A data-driven model inversion approach. *Automatica*, 101, 417 – 430. doi: https://doi.org/10.1016/j.automatica.2018.12.026.

Novara, C., Nicolì, A., and Calafiore, G. (2019). Nonlinear system identification in Sobolev spaces. *arXiv:1911.02930*.

Novara, C., Vincent, T., Hsu, K., Milanese, M., and Poolla, K. (2011). Parametric identification of structured nonlinear systems. *Automatica*, 47(4), 711 – 721. doi: 10.1016/j.automatica.2011.01.063.

Piga, D., Forgione, M., Formentin, S., and Bemporad, A. (2019). Performance-oriented model learning for data-driven mpc design. *IEEE Control Systems Letters*, 3(3), 577–582. doi:10.1109/LCSYS.2019.2913347.

Pukrittayakamee, A., Hagan, M., Raff, L., Bukkapatnam, S.T., and Komanduri, R. (2011). Practical training framework for fitting a function and its derivatives. *IEEE Transactions on Neural Networks*. doi: 10.1109/TNN.2011.2128344.

Qu, Z. (1998). *Robust Control of Nonlinear Uncertain Systems*. Wiley series in nonlinear science.

Salvador, J., de la Peña, D.M., Alamo, T., and Bemporad, A. (2018). Data-based predictive control via direct weight optimization. *IFAC-PapersOnLine*, 51(20), 356 – 361.

Schweppe, F. (1973). *Uncertain dynamic systems*. Prentice-Hall, Englewood Cliffs, NJ.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., B.Delyon, Glorennec, P., Hjalmarsson, H., and Juditsky, A. (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31, 1691–1723.

Sznaier, M., Wenjing, M., Camps, O., and Hwasup, L. (2009). Risk adjusted set membership identification of wiener systems. *IEEE Transactions on Automatic Control*, 54(5), 1147–1152.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Royal. Statist. Soc B.*, 58(1), 267–288.

Traub, J.F., Wasilkowski, G.W., and Woźniakowski, H. (1988). *Information-Based Complexity*. Academic Press, Inc.

Tropp, J. (2006). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3), 1030 –1051. doi: 10.1109/TIT.2005.864420.

Xie, T. and Cao, F. (2011). The errors of simultaneous approximation of multivariate functions by neural networks. *Computers and Mathematics with Applications*, 61, 3146–3152.