Nonlinear system identification in Sobolev spaces

Nonlinear system identification in Sobolev spaces

Carlo Novara[a], Angelo Nicolì[a] and Giuseppe C. Calafiore[a]

[a]Dept. of Electronics and Telecommunications, Politecnico di Torino, Italy

ARTICLE HISTORY
Compiled January 14, 2022

Corresponding Author: Carlo Novara. Email: carlo.novara@polito.it

Nonlinear system identification in Sobolev spaces

Abstract
We consider the problem of approximating an unknown function from experimental data, while approximating at the same time its derivatives. Solving this problem is useful, for instance, in the context of nonlinear system identification, for obtaining models that are more accurate and reliable than the traditional ones based on plain function approximation. Indeed, models identified by accounting for the derivatives can provide improved performance in several endeavours, such as in multi-step prediction, simulation, Nonlinear Model Predictive Control, and control design in general. In this paper, we propose a novel approach based on convex optimization, allowing us to solve the aforementioned identification problem. We develop an optimality analysis, showing that models derived using this approach enjoy suitable optimality properties in Sobolev spaces. The optimality analysis also leads to the derivation of tight uncertainty bounds on the unknown function and its derivatives. We demonstrate the effectiveness of the approach with three numerical examples, concerned with univariate function identification, multi-step prediction of the Chua chaotic circuit, and control of the inverted pendulum.

KEYWORDS
Nonlinear system identification, Set Membership, Sobolev norm approximation.

1. Introduction

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

$$\begin{aligned} y_{k+1} &= f_o\left(x_k\right) + \xi_{k+1} \\ x_k &= \left(y_k, \ldots, y_{k-m_o+1}, u_k, \ldots, u_{k-m_o+1}\right) \end{aligned} \tag{1}$$

where $u_k \in U \subset \mathbb{R}^{n_u}$ is the input, $y_k \in \mathbb{R}^{n_y}$ is the output, $\xi_k \in \Xi \subset \mathbb{R}^{n_\xi}$ is a disturbance and $k \in \mathbb{Z}$ is the discrete time index. The sets U and Ξ are compact with non-empty interior. The *regression function* f_o is supposed to be unknown: the objective of this paper is to obtain 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. Some relevant motivations for considering this problem are given next.

Multi-step prediction and simulation. A standard approach to the identification of system (1) is to adopt a parametrized NARX (Nonlinear Auto Regressive with eXogenous inputs) model structure and to estimate the involved parameters by minimizing the model prediction error; see, e.g., (Ljung, 1999; Sjöberg et al., 1995). A relevant issue is that a model identified using this approach may be accurate when used for one-step ahead prediction but poor when used for multi-step prediction or simulation. This may happen, for example, when the model sampling time is too short; (Goodwin, Yuz, Aguero, & Cea, 2010). In this case, the identified model tends to become a so-called persistent model, where the prediction is close to the current value: $\hat{y}_{k+1} \cong y_k$. Clearly, a nearly persistent model cannot provide a decent performance when used for multi-step prediction

or simulation. In general, the main reason behind these kind of issues is that the model just aims to minimize the one-step prediction error, without really trying to capture the relation between the output and the individual components of x_k and y_{k+1}. An approach that may help overcoming these issues consists in adopting a NOE (Nonlinear Output Error) model structure, in which the involved parameters are estimated by minimizing the model simulation error, see, e.g., (Ljung, 1999; Sjöberg et al., 1995). NOE models are often more accurate than NARX models in multi-step prediction and simulation but require a higher computational burden, since minimization of the simulation error is in general a hard nonlinear and non-convex problem. In any case, also for NOE models there are no guarantees that the relation between the components of x_k and y_{k+1} is correctly captured. The function derivatives express these relations, up to first order. Hence approximating them, together with the system function f_o, is crucial in determining an accurate model for control purposes.

Nonlinear Model Predictive Control (NMPC). NMPC is a widely used technique for controlling complex nonlinear plants, see, e.g., (Findeisen, Allgower, & Biegel, 2007; Grune & Pannek, 2011; Magni, Raimondo, & Allgower, 2009). Data-driven versions of this technique can be found in, e.g., (Manzano, Limon, de la Peñ, & Calliess, 2018; Novara, Formentin, Savaresi, & Milanese, 2016; Novara & Milanese, 2019; Piga, Forgione, Formentin, & Bemporad, 2019; Salvador, de la Peña, Alamo, & Bemporad, 2018). NMPC is based on two main operations: (i) multi-step prediction of the plant behavior, and (ii) synthesis of a control law via on-line optimization, based on the predicted behavior. Clearly, the availability of an accurate multi-step prediction model is of paramount importance in NMPC. In particular, at every time k, given the input and output regressors $(u_{k-1}, \ldots, u_{k-m_o+1})$ and $(y_k, \ldots, y_{k-m_o+1})$, the model should correctly describe the variations of the predicted output $\hat{y}_{k+\tau}$, $\tau \geq 1$, due to variations of the command input sequence $(u_k, \ldots, u_{k+\tau-1})$. The function derivatives describe these variations to first order, and this, again, motivates the need in a control context of approximating the system function f_o together with its derivatives.

Control sensitivity. The above considerations are not limited to NMPC. In general, when estimating a regression model that is to be used for control, e.g., of the type $\hat{y}_{k+1} = \hat{f}(y_k, \ldots, y_{k-m_u+1}, u_k, \ldots, u_{k-m_u+1})$, it is important to capture the sensitivities of the output with respect to the commands $u_k, \ldots, u_{k-m_u+1}$, 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.

Remark 1. Although the following one is an elementary fact, it is perhaps important to remark that a good uniform error bound on a function's values needs not imply a good error bound on the sensitivities (derivatives). Indeed, suppose that we have $\hat{f}(x) = f(x) + e(x)$, where f is the true function, $\hat{f}$ is the identified approximation, and e is an error term. If $\hat{f}$ is approximated in a standard way, we may have that, over a given domain $\mathcal{X}$, $|e(x)| \leq \epsilon$, $\forall x \in \mathcal{X}$, that is a uniform bound ϵ on the absolute approximation error $|\hat{f}(x) - f(x)|$.

The point, however, is that even if ϵ is small, the error on the sensitivity can be arbitrarily large. We have that $\frac{d\hat{f}}{dx} = \frac{df}{dx} + \frac{de}{dx}$, whence $\left|\frac{d\hat{f}}{dx} - \frac{df}{dx}\right| = \left|\frac{de}{dx}\right|$, and indeed it suffices to consider an example with $e(x) = \epsilon \sin(\omega x)$, to see that $|e(x)| \leq \epsilon$ for all x, but $\left|\frac{de}{dx}\right| = \epsilon\omega|\cos(\omega x)|$, thus the error on the sensitivity can be arbitrarily large, for arbitrarily large ω. $\star$

Related literature. The literature appears to be scarce on the topic of approximating from data a function and its derivatives. To the best of the Authors' knowledge, no methods are available in the system identification literature, considering the idea of approximating the system regression function and its derivatives. Regularization and kernel-based methods can be mentioned in this context, see the survey (Pillonetto, Dinuzzo, Chen, De Nicolao, & Ljung, 2014) and the references therein. In these methods, the regression function derivatives are indirectly taken into account by introducing regularization terms in the objective criterion to minimize and/or by using kernel functions characterized by suitable smoothness properties. The few existing methods that directly aim at identifying the regression function and its derivatives are available in the machine learning literature. These methods are based on different classes of approximators, such as radial basis functions (Mai-Duy & Tran-Cong, 2003), neural networks (Avrutskiy, 2018; Pukrittayakamee, Hagan, Raff, Bukkapatnam, & Komanduri, 2011; Xie & Cao, 2011), and deep neural networks (Czarnecki, Osindero, Jaderberg, Swirszcz, & Pascanu, 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 the generalization capabilities. This literature is interesting and effective in showing the potential of techniques relying on derivative identification. However, several issues have still to be addressed in this context: (i) Only a limited number of works carry out a theoretical analysis on the approximation properties of these techniques (Czarnecki et al., 2017; Hornik, Stinchcombe, & White, 1990; Xie & Cao, 2011), and the provided results are often non-constructive, in the sense that they just prove existence of the required approximating function. (ii) The existing methods assume the availability of the function derivative measurements but this may not be true in practical applications. (iii) The existing methods allow for the identification of a model, but they usually do not provide a description of the uncertainty associated with this model and its predictions. (iv) The model learning process is typically based on nonlinear optimization. The identified models may thus be not really optimal since they may correspond to a local minimum of the underlying objective function.

Main Novel Contributions. In this paper, we propose a novel identification approach addressing all the aforementioned issues. The approach provides models approximating not only the system regression function, as done in standard identification methods, but also its derivatives. This allows the identified models to better capture the sensitivity of the output with respect to the input, which in turn leads to improved accuracy and reliability with respect to the standard methods based on plain function approximation (i.e., identified without con-

sidering the derivatives). In summary, the main novelties of the method are as follows:

- To the best of the Authors' knowledge, the approach is the first available in the system identification literature, based on the approximation of the regression function and its derivatives.
- The approach is presented together with a theoretical analysis, showing that the identified models enjoy suitable optimality properties in Sobolev spaces. Roughly speaking, we show that the accuracy of the identified models is close to the best accuracy achievable from the available prior and experimental information.
- The optimality analysis leads to the derivation of tight uncertainty bounds on the unknown function and its derivatives, quantifying the modeling error and the prediction uncertainty.
- The approach uses samples of the regressor x_k, of the function output y_k and of the function derivative outputs. As already mentioned, these latter samples may be not available in a real-world application. Thus, a technique is proposed for estimating the derivative samples from the function input-output data. It is shown that, under standard conditions, the estimate converge to the real values when the number of data becomes large.

Another interesting feature of the approach is that it is completely based on convex optimization. This allows us to avoid the issue of local minima and to propose two identification algorithms that are relatively easy to use in practice.

Three numerical examples are finally presented, concerned with identification of a univariate function, multi-step prediction of the Chua chaotic circuit and control of the inverted pendulum, showing that the approach may provide significantly more accurate and reliable models than the traditional ones based on plain function approximation.

Paper orgnization. In Section 2, the notation used in the paper and some basic notions about functional norms and spaces are introduced. In Section 3, the identification problem of interest is formalized. In Section 4, two methods are discussed for the joint function and derivatives identification problem. The optimality properties of these methods are analyzed in Section 5. Based on this analysis, tight uncertainty bounds are provided in Section 6. In Section 7, an algorithm is proposed for estimating the derivative values, starting from the function input-output values. In Section 8, the main steps of the complete identification procedure are summarized, together with the guidelines about the choice of the involved parameters. Section 9 presents the numerical examples. Conclusions are given in Section 10. All the theorem proofs can be found in the Appendix.

2. Notation and preliminaries

A column vector $x \in \mathbb{R}^{n_x \times 1}$ is denoted by $x = (x_1, \ldots, x_{n_x})$. A row vector $x \in \mathbb{R}^{1 \times n_x}$ is denoted by $x = [x_1, \ldots, x_{n_x}] = (x_1, \ldots, x_{n_x})^\top$, where $^\top$ indi-

cates the transpose. The ℓ_p norm of a vector $x = (x_1, \ldots, x_{n_x})$ 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 $\|\Phi\|_\infty \doteq \max_{i=1,\ldots,m} \sum_{j=1}^{n} |\Phi_{ij}|$.

The $\mathcal{L}_p$ norm of a function with domain $X \subseteq \mathbb{R}^{n_x}$ and codomain in $\mathbb{R}$, is defined as $\|f\|_p \doteq \left[\int_X |f(x)|^p dx\right]^{\frac{1}{p}}$, for $p \in (1, \infty)$, and as $\|f\|_\infty \doteq \operatorname{ess\,sup}_{x\in X} |f(x)|$ 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 by 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 $f^{(i)} \in \mathcal{L}_p(X)$. That is, $\mathcal{S}_{1p}(X) \doteq \left\{f \in \mathcal{L}_p(X) : f^{(i)} \in \mathcal{L}_p(X),\ i = 0, \ldots, n_x\right\}$.*

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, which is more relevant from a practical point of view. The generalization to the case of weak derivatives is in any case straightforward.

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:

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

where $\tilde{x}_k \in X$ are the measurements of f_o argument, $\tilde{z}_{0,k} \equiv \tilde{z}_k$ are the measurements of f_o output, and $\tilde{z}_{i,k}$, $i > 0$, are the measurements of $f_o^{(i)} \doteq \partial f_o/\partial x_i$ output. If multiple datasets $D^{[j]} = \{\tilde{x}_k^{[j]}, \{\tilde{z}_k^{[j],i}\}_{i=0}^{n_x}\}_{k=1}^{L^{[j]}}$, $j = 1, \ldots, N_E$ are available, collected in different experiments, then the overall dataset D is given by their union: $D = \cup_{j=1}^{N_E} D^{[j]}$.

The data (2) can be described by

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

where $d_{i,k}$ are noises accounting for input/output disturbances, and $d_{0,k} \equiv d_k$.

If the data are generated by the system (1), we have that $\tilde{z}_{0,k} \equiv \tilde{z}_k = \tilde{y}_{k+1}$, where $\tilde{y}_k$ indicates the measured value of y_k, and the noise terms account for the disturbance ξ_k 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. In Section 7, an algorithm will be presented for estimating the derivative output samples $\tilde{z}_{i,k}$, $i > 0$, from the input-output function samples $\tilde{x}_k$ and $\tilde{z}_k$. In the reminder of this paper, we will thus assume that these samples are available, because they have been either measured or estimated. All the results that will be presented will hold true for both these cases.

Now, assume that the noise sequences $d_i = (d_{i,1}, \ldots, d_{i,L})$ 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 will be provided in Section 6 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 Sobolel 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., (Novara, Vincent, Hsu, Milanese, & Poolla, 2011; Sjöberg et al., 1995). 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 9.2 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.

The problem considered in this paper is stated as follows.

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.

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. We suppose that the derivative output samples $\tilde{z}_{i,k}$, $i > 0$ are available. In Section 7, we will show how these derivative samples can be estimated from the input-output function samples $\tilde{x}_k$ and $\tilde{z}_k$.

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

$$\tilde{z}_i \doteq \begin{bmatrix} \tilde{z}_{i,1} \\ \vdots \\ \tilde{z}_{i,L} \end{bmatrix}, \; \Phi_i \doteq \begin{bmatrix} \phi_1^{(i)}(\tilde{x}_1) & \cdots & \phi_N^{(i)}(\tilde{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1^{(i)}(\tilde{x}_L) & \cdots & \phi_N^{(i)}(\tilde{x}_L) \end{bmatrix}. \tag{7}$$

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

the following convex optimization problem:

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

$$\text{s.t. } \|\tilde{z}_i - \Phi_i \alpha\|_q \leq \mu_i, \; i = 0, \ldots, n_x, \tag{9}$$

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., (Donoho, Elad, & Temlyakov, 2006; Fuchs, 2005; Tibshirani, 1996; Tropp, 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:

$$a = \arg\min_{\alpha \in \mathbb{R}^N} \sum_{i=0}^{n_x} \lambda_i \|\tilde{z}_i - \Phi_i \alpha\|_q^2 + \Lambda \|\alpha\|_r \tag{10}$$

where the integers r, q indicate suitable vector norms, and $\lambda_i \geq 0, \Lambda \geq 0$ are user-defined weights.

Problem (10) is aimed at minimizing a tradeoff between the model fitting error on the identification data and a regularization term. For $r = 1$ and $\lambda_i = 0$, $i > 0$, (10) is a Lasso problem, see, e.g., (Tibshirani, 1996); for $r = 2$ and $\lambda_i = 0$, $i > 0$, 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.

Remark 2. 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. $\star$

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 & Vicino, 1991), (Milanese, Norton, Lahanier, & Walter, 1996), (Schweppe, 1973), (Chen & Gu, 2000), (Milanese & Novara, 2011), (Sznaier, Wenjing, Camps, & Hwasup, 2009), we show that such approximations enjoy certain optimality properties in Sobolev spaces. Two cases are covered: in the first one, we suppose that the true function f_o belongs to a Sobolev space $\mathcal{S}_{1p}$; in the second one, we make an additional assumption, regarding the Lipschitz continuity of the derivatives of the function $f_o - \hat{f}$, which allows us to prove stronger optimality properties of the approximations w.r.t. the first case. The analysis and results developed here are extensions to Sobolev spaces of those regarding approximation in $\mathcal{L}_p$ spaces presented in (Milanese & Novara, 2004, 2011).

Before proceeding with the optimality analysis, it is important to observe that any system identification method is based on some prior assumptions. For example, the well-known Prediction Error method (Ljung, 1999) assumes given statistical properties of the noise (e.g., mean, variance, covariance, uncorrelation, type of distribution, etc.) and a parametric structure of the model to identify. If these assumptions are true, then it is possible to obtain valid theoretical guarantees about the identified model, otherwise these guarantees are not reliable. In our approach, we do not assume statistical properties of the noise and parametric structures of the true system. Following a Set Membership setting, we suppose that the noise is unknown but bounded and that the true system regression function is characterized by suitable regularity properties, i.e., that this function belongs to the Sobolev space $\mathcal{S}_{1p}(X)$.

Prior Assumptions on the noise: $\|d_i\|_q \leq \mu_i$, $i = 0, \ldots, n_x$.

Prior Assumptions on the regression function: $f_o \in \mathcal{S}_{1p}(X)$.

Under these assumptions, we are going to derive optimality guarantees about the accuracy of the identified models and tight bounds on the unknown function f_o and its derivatives.
It must be remarked that it is not possible to be guaranteed that the considered prior assumptions are true. What can be actually done is to validate the assumptions using the available experimental data (Popper, 1969). In the following, we will provide conditions for prior assumption validation and procedures for estimating the involved parameters.

5.1. Optimality analysis in Sobolev spaces

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 assumptions given by the inclusion $f_o \in \mathcal{S}_{1p}(X)$ and the noise bounds $\|d_i\|_q \leq \mu_i$. It follows that $f_o \in \mathrm{FFS}_{\mathcal{S}}$, where $\mathrm{FFS}_{\mathcal{S}}$ is the so-called Feasible Function Set,

defined below.

Definition 2. *The Feasible Function Set* $\mathrm{FFS}_{\mathcal{S}}$ *is defined as*

$$\mathrm{FFS}_{\mathcal{S}} \doteq \{f \in \mathcal{S}_{1p}(X) : ||\tilde{z}_i - f^{(i)}(\tilde{x})||_q \leq \mu_i, i = 0, \ldots, n_x\}$$

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 $\mathrm{FFS}_{\mathcal{S}} \neq \emptyset$), we say that the assumptions are validated. Otherwise (i.e., if $\mathrm{FFS}_{\mathcal{S}} = \emptyset$), we say that the assumptions are falsified; see (Chen & Gu, 2000; Milanese et al., 1996).

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

As discussed above, assumption validation is an important step of the model identification process since, without validation, all the theoretical properties that can be derived are not reliable. However, checking the non-emptiness of the Feasible Function Set is in general not trivial. The following theorem gives a sufficient condition for $\mathrm{FFS}_{\mathcal{S}}$ to be non-empty. It can be noted that the condition is easily verifiable in practice.

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

Proof. See the Appendix. □

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, $f_o \in \mathrm{FFS}_{\mathcal{S}}$. 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* $WE(\hat{g}, \mathrm{FFS}_{\mathcal{S}}) \doteq \sup_{f \in \mathrm{FFS}_{\mathcal{S}}} \|f - \hat{g}\|_{\mathcal{S}p}$, *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* $\mathrm{FFS}_{\mathcal{S}}$*-optimal if* $WE(f_{op}, \mathrm{FFS}_{\mathcal{S}}) = \inf_{\hat{g}} WE(\hat{g}, \mathrm{FFS}_{\mathcal{S}}) \doteq \mathcal{R}(\mathrm{FFS}_{\mathcal{S}})$, *where* $\mathcal{R}(\mathrm{FFS}_{\mathcal{S}})$ *is called the* radius of informa-

tion *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, Wasilkowski, & Woźniakowski, 1988), (Milanese et al., 1996).

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

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* $\mathrm{FFS}_{\mathcal{S}}$*-almost-optimal.*

Proof. See the Appendix. □

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).

5.2. Optimality analysis with Lipschitz information

As discussed in Section 5.1, the function f_o and its derivatives are unknown, while instead we have available the experimental information given by (2) and (3), and the prior assumptions given by the inclusion $f_o \in \mathcal{S}_{1p}(X)$ and the noise bounds. In this section, we make an additional assumption on the Lipschitz continuity of the derivatives of the so-called residue function $f_o - \hat{f}$. This allows us to prove stronger optimality properties with respect to those discussed in Section 5.1.

The residue function is defined as $\Delta(x) \doteq f_o(x) - \hat{f}(x)$. We assume that this function and its derivatives are Lipschitz continuous. That is, for given Lipschitz constants $\gamma_i < \infty$, $i = 0, \ldots, n_x$, $\Delta^{(i)} \in \mathcal{L}(\gamma_i, X)$, where

$$\mathcal{L}(\eta, X) \doteq \{f \in \mathcal{S}_{1p}(X) : |f(x) - f(w)| \leq \eta \|x - w\|_\infty, \forall x, w \in X\}.$$

This assumption is reasonable, since we already know that $\Delta \in \mathcal{S}_{1p}(X)$, which implies that Δ is Lipschitz continuous and its derivatives are continuous (a slightly weaker condition with respect to Lipschitz continuity). The constants γ_i can be estimated from the available data by means of the procedure presented

at the end of this section.

Under the Lipschitz condition, we have that $f_o \in \mathrm{FFS}_{\mathcal{L}}$, where $\mathrm{FFS}_{\mathcal{L}}$ is the following Feasible Function Set.

Definition 7. *We let*

$$\mathrm{FFS}_{\mathcal{L}} \doteq \{f \in \mathcal{S}_{1p}(X) : f^{(i)} - \hat{f}^{(i)} \in \mathcal{L}(\gamma_i, X), ||\tilde{z}_i - f^{(i)}(\tilde{x})||_q \leq \mu_i, i = 0, \ldots, n_x\}$$

where $f^{(i)}(\tilde{x}) \doteq (f^{(i)}(\tilde{x}_1), \ldots, f^{(i)}(\tilde{x}_L))$.

$\mathrm{FFS}_{\mathcal{L}}$ is the set of all functions consistent with the prior assumptions and the available data. As discussed above, assumption validation is an important step of the model identification process since, without validation, all the theoretical properties that can be derived are not reliable. Recalling Definition 3, a result is now presented, giving sufficient conditions for assumption validation.

Theorem 3. $\mathrm{FFS}_{\mathcal{L}} \neq \emptyset$ *if the optimization problem (8)-(9) is feasible.*

Proof. See the Appendix. □

To see how the assumption about the Lipschitz continuity of the function derivatives helps to obtain stronger optimality properties, consider Definitions 2 and 7. These definitions imply that $\mathrm{FFS}_{\mathcal{L}} \subseteq \mathrm{FFS}_{\mathcal{S}}$ and, consequently, $\mathcal{R}(\mathrm{FFS}_{\mathcal{L}}) \leq \mathcal{R}(\mathrm{FFS}_{\mathcal{S}})$. This inequality shows that the Lipschitz continuity assumption yields a reduction of the worts-case identification error.

The following result gives sufficient conditions, under which an approximation is almost-optimal, when the Feasible System Set is $\mathrm{FFS}_{\mathcal{L}}$.

Theorem 4. *Let the assumptions of Theorem 2 hold and the functions* $\Delta^{(i)}$, $i = 1, \ldots, n_x$, *be Lipschitz continuous. Then, the approximation* $\hat{f}$ *is* $\mathrm{FFS}_{\mathcal{L}}$*-almost-optimal.*

Proof. See the Appendix. □

This section is concluded with a procedure for estimating the constants γ_i from the available data. The procedure is the following:

(1) Let $\Delta\tilde{z}_{i,k} \doteq \tilde{z}_{i,k} - \hat{f}^{(i)}(\tilde{x}_k)$. The values $\tilde{z}_{i,k}$, $k = 1, \ldots, L$, are either known/measured or estimated from the data $\{\tilde{x}_k, \tilde{z}_k\}_{k=1}^{L}$, using Algorithm 1 presented next in Section 7.

(2) Let $\Delta\tilde{z}_{ij,k}$ be the samples of the derivative with respect to the jth variable of $\Delta^{(i)}$. The values $\Delta\tilde{z}_{ij,k}$, $k = 1, \ldots, L$, $i, j = 1, \ldots, n_x$, are estimated from the data $\left\{\tilde{x}_k, \tilde{z}_{i,k} - \hat{f}^{(i)}(\tilde{x}_k)\right\}_{k=1}^{L}$, using Algorithm 1. Note that the estimation of the $\Delta\tilde{z}_{ij,k}$'s requires the function f_o to be locally twice differentiable at the points $\tilde{x}_k$, $k = 1, \ldots, L$.

(3) Estimate the Lipschitz constants γ_i, $i = 0, \ldots, n_x$, as

$$\begin{aligned} \gamma_0 &= \nu \max_{k=1,\ldots,L} \|(\Delta\tilde{z}_{1,k}, \ldots, \Delta\tilde{z}_{n_x,k})\|_\infty \\ \gamma_i &= \nu \max_{k=1,\ldots,L} \|(\Delta\tilde{z}_{i1,k}, \ldots, \Delta\tilde{z}_{in_x,k})\|_\infty \end{aligned} \tag{11}$$

where $\nu \gtrsim 1$ is a coefficient introduced to guarantee a desired safety level.

This procedure is based on the observation that the Lipschitz constant of a differentiable function is an upper bound of the function's gradient norm, which gives the motivation for (11).

6. Uncertainty bounds

In this section, we derive tight uncertainty bounds for the unknown function f_o and its derivatives $f_o^{(i)}$, $i = 1, \ldots, n_x$. These bounds allow us to 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 & Kokotovic, 1996), (Qu, 1998), prediction interval evaluation (Milanese & Novara, 2005), and fault detection (Novara, 2016). Based on the uncertainty bounds, we present an algorithm allowing us to estimate the noise bounds μ_i. The result presented here regarding the uncertainty bound derivation is an extension of the one in (Milanese & Novara, 2004, 2011; Novara, 2016) to the case where the bounds are derived not only for a function but also for its first-order derivatives.

Under the Lipschitz assumption $\Delta^{(i)} \in \mathcal{L}(\gamma_i, X)$, we can define the following functions:

$$\begin{aligned} \overline{\Delta}_i(x) &\doteq \min_{k=1,\ldots,L} \left(\overline{h}_{i,k} + \gamma_i \|x - \tilde{x}_k\|_\infty\right) \\ \underline{\Delta}_i(x) &\doteq \max_{k=1,\ldots,L} \left(\underline{h}_{i,k} - \gamma_i \|x - \tilde{x}_k\|_\infty\right) \end{aligned} \tag{12}$$

where $\overline{h}_{i,k} = \tilde{z}_{i,k} - \hat{f}^{(i)}(\tilde{x}_k) + \mu_i$, $\underline{h}_{i,k} = \tilde{z}_{i,k} - \hat{f}^{(i)}(\tilde{x}_k) - \mu_i$ and $i = 0, \ldots, n_x$.

A result is now presented, providing tight uncertainty bounds in closed form for the unknown function f_o and its derivatives $f_o^{(i)}$. The result holds in the case where the noise is bounded in ℓ_∞ norm. The case where the noise is bounded in ℓ_2 norm is discussed afterwards.

Theorem 5. *Let the assumptions of Theorem 2 hold and $q = \infty$ in the noise bounds $\|d_i\|_q \leq \mu_i$. Then, $\underline{f}_i(x) \leq f_o^{(i)}(x) \leq \overline{f}_i(x)$, where*

$$\begin{aligned} \overline{f}_i(x) &= \hat{f}^{(i)}(x) + \min\left(\bar{\gamma}, \overline{\Delta}_i(x)\right) \\ \underline{f}_i(x) &= \hat{f}^{(i)}(x) + \max\left(-\bar{\gamma}, +\underline{\Delta}_i(x)\right) \end{aligned} \tag{13}$$

and $\bar{\gamma} \doteq \infty$ *if* $i = 0$ *or* $\bar{\gamma} \doteq \gamma_0$ *otherwise. Moreover,*

$$\begin{aligned} \overline{f}_i(x) &\doteq \sup_{f \in \mathcal{F}_i} f(x) \\ \underline{f}_i(x) &\doteq \inf_{f \in \mathcal{F}_i} f(x) \end{aligned} \tag{14}$$

where $\mathcal{F}_i \doteq \{f : f - \hat{f}^{(i)} \in \mathcal{L}(\gamma_i, X), ||\tilde{z}_i - f(\tilde{x})||_\infty \leq \mu_i\}$.

Proof. See the Appendix. □

This theorem shows that, for a given $i \in \{0, \ldots, n_x\}$, $\overline{f}_i$ and $\underline{f}_i$ are the tightest upper and lower bounds of $f_o^{(i)}$ that can be defined on the basis of the information available about $f_o^{(i)}$, summarized by the function set $\mathcal{F}_i$. This result is important since it shows that the bounds $\overline{f}_i$ and $\underline{f}_i$ are tight. Examples of these bounds are reported in Figure 3 below. Note that improved bounds on $f_o^{(i)}$ could be formally defined under the assumption $f \in \text{FFS}_{\mathcal{L}}$ instead of $f \in \mathcal{F}_i$. However, the evaluation of such bounds would be hard from a computational point of view. On the contrary, the bounds (13) are written in closed form and are simple to evaluate.

Remark 3. It can be proven that the function f_c defined as $f_c(x) \doteq \frac{1}{2}\left(\overline{f}_0(x) + \underline{f}_0(x)\right)$ is an optimal approximation of f_o in any $\mathcal{L}_p$ space (Milanese & Novara, 2004). However, f_c is not an optimal approximation of f_o in a Sobolev space. Indeed, the derivatives of f_c are discontinuous and thus are not completely appropriate for approximating the derivatives of f_o, which instead are continuous. ⋆

In the case where the noise is bounded in ℓ_2 norm (i.e., $q = 2$ in the noise bounds $\|d_i\|_q \leq \mu_i$, Theorem 5 cannot be applied as it is, since the ℓ_2 norm bound on the sequence gives no information on how the individual elements $d_{i,k}$ are bounded. In order to overcome this issue, some additional assumption has to be made on the element-wise boundedness of the noise sequence d_i. Suppose that the estimates $\hat{f}^{(i)}$ obtained from some of the two identification methods in Section 4 are sufficiently accurate approximations of the functions $f_o^{(i)}$: $\hat{f}^{(i)}(\tilde{x}_k) \cong f_o^{(i)}(\tilde{x}_k)$. It follows that $d_{i,k} = \tilde{z}_{i,k} - f_o^{(i)}(\tilde{x}_k) \cong \tilde{z}_{i,k} - \hat{f}^{(i)}(\tilde{x}_k) \doteq \delta_{i,k}$. It is then natural to consider the following relative-plus-absolute error bound:

$$|d_{i,k}| \leq \zeta_{i,k} \doteq \zeta_{R,i}\,|\delta_{i,k}| + \zeta_{A,i}, \quad k = 1, \ldots, L \tag{15}$$

where the term $\zeta_{R,i}\,|\delta_{i,k}|$ accounts for the fact that $d_{i,k} \cong \delta_{i,k}$ and $\zeta_{A,i}$ accounts for the fact that $d_{i,k}$ and $\delta_{i,k}$ are not exactly equal. The parameters $\zeta_{R,i}, \zeta_{A,i} \geq 0$ have to be taken such that $\zeta_{R,i}\mu_i + \zeta_{A,i}\sqrt{L} \leq \mu_i$. Indeed, if this inequality is satisfied, (15) is consistent with $\|d_i\|_q \leq \mu_i$, since $\|d_i\|_2 \leq \zeta_{R,i}\mu_i + \zeta_{A,i}\sqrt{L} \leq \mu_i$. Following this indication, $\zeta_{R,i}$ and $\zeta_{A,i}$ can be chosen by means of the procedure presented at the end of this section.

Assuming the bound (15), Theorem 5 holds, where the functions $\overline{\Delta}_i$ and $\underline{\Delta}_i$ in (13) are defined as in (12), with $\mu_i \to \zeta_{i,k}$.

Now, a procedure for estimating the noise bounds μ_i in (4) is proposed, based on the optimal function bounds given in Theorem 5. For a given i, consider the case where $\hat{f}^{(i)}(x) = 0$, $\forall x \in X$. Suppose that the Lipschitz constant γ_i of the function $\Delta^{(i)} \doteq f_o^{(i)} - \hat{f}^{(i)} = f_o^{(i)}$ has been estimated by means of Algorithm 1 in Section 7. According to Theorem 5, for some suitable $\mu_i \geq 0$, the functions $\overline{f}_i$ and $\underline{f}_i$ in (13) are upper and lower bounds of the unknown function $f_o^{(i)}$. Clearly, it must hold that $\overline{f}_i(x) > \underline{f}_i(x)$, $\forall x \in X$. The following procedure provides an estimate of μ_i such that this inequality is met on the measured data.

(1) Let $\hat{f}^{(i)}(x) = 0$, $\forall x \in X$.

(2) Solve the following optimization problem:

$$\begin{aligned} \underline{\mu}_i &= \min_{\mu_i \geq 0} \mu_i \\ \text{s.t. } & \overline{f}_i(\tilde{x}_k) - \underline{f}_i(\tilde{x}_k) > 0, \ k = 1, \ldots, L. \end{aligned} \tag{16}$$

(3) Estimate the noise bound as $\hat{\mu}_i = \nu \underline{\mu}_i$, where $\nu \gtrsim 1$ is a coefficient introduced to guarantee a desired safety level.

The optimization problem (16) can be easily solved since the decision variable $\underline{\mu}_i$ is scalar and the number of constraints is finite. Notice that the difference $\overline{f}_i(\tilde{x}_k) - \underline{f}_i(\tilde{x}_k)$ does not depend on $\hat{f}^{(i)}$. It follows that the above procedure uses only data to estimate the noise bounds, and no approximations of the unknown function are required.

As discussed at the beginning of Section 5, any system identification method is based on suitable prior assumptions. In this paper, following a Set Membership philosophy, we assume that the noise is unknown but bounded, and that the regression function of the true system belongs to a suitable Sobolev space. Considering also the Lipschitz property of the function and its derivatives, these assumptions are summarized by the parameters μ_i and γ_i. Procedures for estimating the values of these parameters have been provided above. It must be remarked that it is not possible to obtain theoretical guarantees about the obtained values without other additional assumptions. According to (Popper, 1969), what can be actually done is to validate these values using the available data. This validation can be carried out according to the following procedure.

(1) Estimate the parameters γ_i using the procedure of Section 5.2.
(2) Based on the obtained γ_i values, estimate the parameters μ_i using the procedure of this section.
(3) Apply Theorem 3.

7. Estimation of the derivative values

In practical situations, only the output of the function that describes the system of interest is usually measured, while the outputs of its derivatives are not. In this section, we propose an algorithm for estimating the derivative output samples $\tilde{z}_{i,k}$, $i > 0$, from the input-output function samples $\tilde{x}_k$ and $\tilde{z}_k$.

Suppose that the dataset $D_0 = \{\tilde{x}_k, \tilde{z}_k\}_{k=1}^{L}$ is available. The algorithm for estimating the derivative output samples $\tilde{z}_{i,k}$, $i > 0$, is the following.

Algorithm 1. *For $k = 1, \ldots, L$:*

(1) Define the set of indexes

$$\Upsilon_{\rho k} \doteq \{j \in \{1, \ldots, L\} : \|\tilde{x}_j - \tilde{x}_k\|_2 \leq \rho\}$$

where $\rho > 0$ is a user-defined radius.

(2) Define the following quantities:

$$\tilde{z}_{\rho k} \doteq \begin{bmatrix} \tilde{z}_{j_1} - \tilde{z}_k \\ \vdots \\ \tilde{z}_{j_M} - \tilde{z}_k \end{bmatrix}, \ \Phi_{\rho k} \doteq \begin{bmatrix} \tilde{x}_{j_1}^\top - \tilde{x}_k^\top \\ \vdots \\ \tilde{x}_{j_M}^\top - \tilde{x}_k^\top \end{bmatrix}$$

where $\{j_1, \ldots, j_M\} = \Upsilon_{\rho k}$, $M = \mathrm{card}\Upsilon_{\rho k}$ and card *denotes the set cardinality.*

(3) Compute

$$g_k = \arg\min_{\mathfrak{g} \in \mathbb{R}^{n_x}} \frac{1}{M} \|\tilde{z}_{\rho k} - \Phi_{\rho k}\mathfrak{g}\|_2^2. \tag{17}$$

(4) Estimate the derivative output samples as $\tilde{z}_{i,k} = g_{ki}$, $k = 1, \ldots, L$, $i = 1, \ldots, n_x$, where g_{ki} are the components of g_k.

(5) In the case where the data are affected by a relevant noise and/or the data set is not sufficiently large, the estimated gradient sequence $\{g_k\}_{k=1}^{L}$ can be smoothed by means of a suitable anti-causal discrete-time filter. □

The idea behind the algorithm is to identify a local linear model at each point $\tilde{x}_k$ (steps 1-3). The gradient of f_o is then estimated by taking the gradient of this local model, whose coefficients are indeed the gradient components (step 4). Letting the symbol $\succ$ stand for "positive-definite", the following result provides a bound on the gradient estimation error.

Theorem 6. *Assume that:*

(i) The derivatives $f_o^{(i)}$, $i = 1, \ldots, n_x$, are Lipschitz continuous on X.

(ii) For any $\rho > 0$, a $M_0 > 0$ exists such that $\frac{1}{M}\Phi_{\rho k}^\top \Phi_{\rho k} \succ 0$, $\forall M \geq M_0$.

Then, for any $\epsilon > 0$, some $M_0 > 0$ and $\rho > 0$ exist such that the gradient estimation error is bounded as

$$\|\nabla f_o(\tilde{x}_k) - g_k\|_q \leq 2\|\Phi_{\rho k}^{\dagger}\|_q \mu_0 + \epsilon, \quad \forall M \geq M_0 \tag{18}$$

where $\Phi_{\rho k}^{\dagger} \doteq (\Phi_{\rho k}^{\top}\Phi_{\rho k})^{-1}\Phi_{\rho k}^{\top}$ is the pseudo-inverse matrix of $\Phi_{\rho k}$ and $q \in \{2, \infty\}$.

Proof. See the Appendix. □

This theorem can be interpreted as follows. Two main conditions are sufficient for obtaining a bound on the gradient estimation error. The first one (assumption (i) in the theorem) is Lipschitz continuity of the derivatives $f_o^{(i)}$, $i = 1, \ldots, n_x$. This assumption is reasonable, since we already know that $f_o \in \mathcal{S}_{1p}(X)$, which implies that $f_o^{(i)}$, $i = 1, \ldots, n_x$, are continuous (a slightly weaker condition with respect to Lipschitz continuity). The second one (assumption (ii)) is a standard persistence of excitation condition (Ljung, 1999; Novara et al., 2011). The next result shows that, under these two assumptions and some further technical conditions, the gradient estimate converges to its true value as $\rho \to 0$ and $M \to \infty$.

Theorem 7. *Let the assumptions of Theorem 6 be true. Let also the following limits hold:*

$$\lim_{\rho \to 0} \lim_{M \to \infty} \frac{1}{M} D_k^{\top} D_k = \sigma_D^2 \tag{19}$$

$$\lim_{\rho \to 0} \lim_{M \to \infty} \frac{1}{M} D_k^{\top} \Phi_{\rho k} = 0 \tag{20}$$

where $D_k \doteq (d_{j_1} - d_k, \ldots, d_{j_M} - d_k)$ and $0 \leq \sigma_D^2 < \infty$. Then,

$$\lim_{\rho \to 0} \lim_{M \to \infty} \|\nabla f_o(\tilde{x}_k) - g_k\|_q = 0.$$

Proof. See the Appendix. □

This theorem shows that, in order to ensure convergence of the estimate to the true gradient, the limits (19) and (20) must hold (besides the basic assumptions of Theorem 6). The limit (19) means convergence of the sample noise variance. The limit (20) implies sample uncorrelation between the noise and the regressor. Both these limits (in their statistical version) represent standard assumptions in the literature on system identification, see, e.g., (Ljung, 1999).

8. Summary of the identification procedure

In this section, the main steps of the complete identification procedure are summarized. Guidelines about the choice of the involved parameters are also given.

Suppose that the dataset (2) is available. The samples of the regression function derivatives are either measured or estimated using Algorithm 1. In this algorithm,

the radius ρ can be chosen by means a simple trial and error procedure. The main steps of the identification procedure for Methods 1 and 2 are as follows.

(1) Estimate the parameters γ_i using the procedure of Section 5.2. Based on the obtained γ_i values, estimate the parameters μ_i using the procedure of Section 6. Apply Theorem 3 to validate these estimates.
(2) Choose the basis functions ϕ_j according to the indications given in Section 3.
(3) Choose the integer r on the basis of the desired regularization properties, and the integer q on the basis of the adopted assumptions on the noise.
(4) Only for Method 2. Choose the weights λ_i and Λ according to the desired trade-off between model fitting accuracy and regularity.
(5) Apply Method 1 or 2.

The result of this procedure is the coefficient vector $a = (a_1, \ldots, a_N)$. Given this vector, the model output is evaluated according to equation (5). The model derivative outputs are obtained using (6). The uncertainty bounds are evaluated by means of (13).

Remark 4. In real applications, the regressor order m_o in (1) is often not known and needs to be chosen. The problem of regressor order choice is common to all identification methods. This problem is certainly relevant, but it is a theoretically unsolved problem in almost all methods for nonlinear systems. Hence, its solution is out of the scope of the present paper. We observe, however, that in our identification method, the regressor order can be estimated via a standard validation approach, by running multiple experiments with different orders and verifying the ensuing results on validation data. $\star$

9. Numerical examples

Three numerical examples are presented in this section. The first one is concerned with identification of a univariate function. Although simple, this example is useful to illustrate the proposed methodology and highlight the importance of identifying the derivative of the unknown function. The second one is about multi-step prediction of the Chua chaotic circuit. The third one shows an application of the proposed method in the context of nonlinear model predictive control of a pendulum. In the examples, model identification and parameter choice have been carried out according to the procedure described in Section 8.

9.1. Example: univariate function approximation

The following univariate function is considered in this example:

$$f_o(x) = \sin(1.1x)$$

Table 1. RMSE errors on the validation set.

Estimations	RMSE	RMSE$^{(1)}$
Model 1	2.54e-02	5.86e-02
Model 2	1.19e-02	2.57e-02

where $x \in \mathbb{R}$ and $f_o : \mathbb{R} \to \mathbb{R}$.

The function and its derivative were evaluated in $L = 100$ linearly equally spaced points in the domain $X = [-2, 3]$. A normally distributed noise with zero mean and standard deviation (std) equal to 0.05 was added to the function samples and its derivative samples (the noise was truncated at 3-std). Hence, a noise-corrupted identification dataset of the form (2) was obtained. A validation dataset of length $L = 1000$ was similarly obtained in the same domain X. This set consists of noise-free data, in order to compare the output of the models that will be identified with the true function.

A model of the form (5) was considered, with a basis function set composed of univariate monomials up to degree 5. Two models were identified from the identification dataset:

- *Model 1.* Function values used for model identification, function derivative values not used. The coefficients a_j in (5) were identified by Method 2, with $q = 2$, $r = 1$, $\lambda_0 = 1$, $\lambda_1 = 0$, and $\Lambda = 1$.
- *Model 2.* Both function and derivative values used for model identification. The derivative values were computed analytically. The coefficients a_j in (5) were identified by Method 2, with $q = 2$, $r = 1$, $\lambda_0 = 1$, $\lambda_1 = 2$, and $\Lambda = 1$.

The results obtained by the two models on the validation dataset are summarized in Table 1, where the obtained Root Mean Square Errors are reported. In the Table, RMSE is the error between the true function f_o and the model $\hat{f}$; RMSE$^{(1)}$ is the error between the true function derivative $f_o^{(1)}$ and the model derivative $\hat{f}^{(1)}$. The upper plot in Figure 1 shows the comparison between the true function values and the outputs of the identified models. The lower plot in Figure 1 shows the comparison between the true derivative values and the outputs of the model derivatives. In Figure 2, the Model 2 uncertainty bounds, computed according to (13), are reported.

From these results, we can conclude that the model identified using the derivative values (Model 2) provides a more accurate approximation of the true function derivative with respect to the model identified not using the derivative values (Model 1). What is quite interesting is that Model 2 provides also a better approximation of the true function itself.

9.2. Example: multi-step prediction for the Chua chaotic circuit

The Chua system is a simple electronic circuit showing a chaotic behavior, see (Chua, Komuro, & Matsumoto, 1986). It is composed of two capacitors, an in-

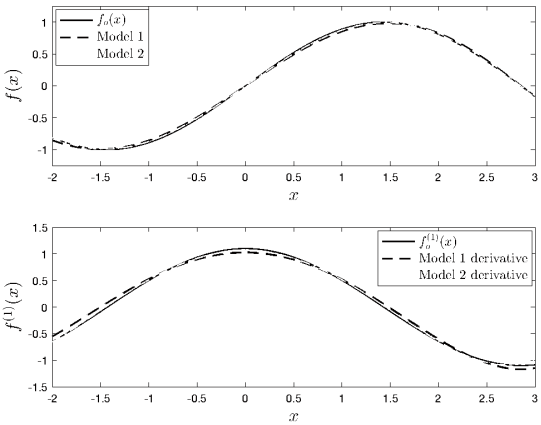

Figure 1. Validation set. Upper plot: comparison between true function and model outputs. Lower plot: comparison between true derivative and model derivatives.

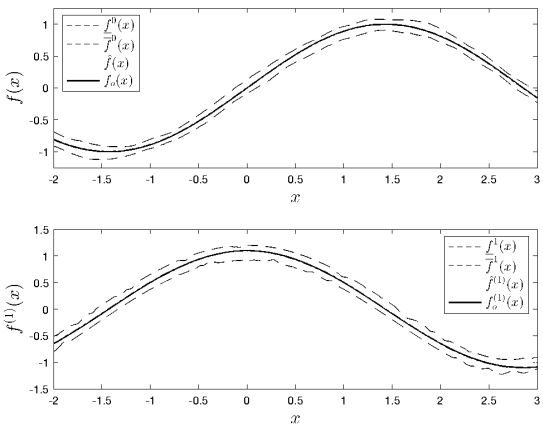

Figure 2. Validation set. True function, Model 2 output, derivative and related uncertainty bounds.

ductor, a locally active resistor and a nonlinear resistor. The continuous-time state equations of the Chua circuit considered here are the following:

$$\begin{aligned} \dot{x}_1 &= \alpha(x_2 - x_1 - \rho(x_1)) \\ \dot{x}_2 &= x_1 - x_2 + x_3 + u + \xi^c \\ \dot{x}_3 &= -\beta x_2 - R x_3 \\ y &= x_1 \end{aligned} \tag{21}$$

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 $\alpha, \beta, R \in \mathbb{R}$ 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 (21), discretized via the forward Euler method, can be written in the following input-output regression form:

$$\begin{aligned} y_k &= b_1 y_{k-1} + b_2 y_{k-2} + b_3 y_{k-3} \\ &\quad + b_4 \rho(y_{k-1}) + b_5 \rho(y_{k-2}) + b_6 \rho(y_{k-3}) \\ &\quad + b_7 u_{k-2} + b_8 u_{k-3} + \xi_k \end{aligned} \tag{22}$$

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

The system (21) 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 and std 0.05, truncated at 3-std. Two simulations of duration 60 s were carried out and, correspondingly, two sets 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. Then, 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

$$\begin{aligned} y_{k+1} &= \hat{f}\left(x_k\right) \\ x_k &= (y_k, y_{k-1}, y_{k-2}, u_{k-1}, u_{k-2}) \end{aligned} \tag{23}$$

  where $\hat{f}$ is of the form (5). A basis function set composed of multivariate

monomials has been used, defined as

$$\{\phi_j\}_{j=1}^{N} = \{\prod_{l=1}^{n_x} x_{l,k}^{\alpha_l} : \alpha_l = 0, 1; l = 1, \ldots, n_x\} \tag{24}$$

where $x_{l,k}$ is the lth component of x_k 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 $\Lambda = 50$. Note that, with these parameter values, Method 2 corresponds to the classical Lasso algorithm.

- *One-step predictor identified using the true derivative values (P1_D).* The predictor P1_D is of the form (23). The basis functions are the same as those used in (23). The true derivative values computed from (22) 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 (23). The basis functions are the same as those used in (23). The derivative values estimated by Algorithm 1 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 $q = 2$, $r = 1$, $\lambda_0 = 1$, $\lambda_i = 200$, $i > 0$, $\Lambda = 50$, and $\rho = 0.4$.
- *Direct multi-step predictor identified not using any derivative information (PK_NOD).* The predictor PK_NOD is given by

$$\begin{aligned} y_{k+\tau} &= \hat{f}\left(x_k\right) \\ x_k &= \left(y_k, y_{k-1}, y_{k-2}, u_{k+\tau-2}, u_{k+\tau-3}, \ldots, u_{k-2}\right) \end{aligned} \tag{25}$$

  where $\hat{f}$ is of the form (5) and $\tau \in \{3, 5, 7\}$. The basis function set is defined as in (24), with $n_x = 4 + \tau$. This set consists of $N = 2^{n_x}$ 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 $\Lambda = 50$. Note that, with these parameter values, Method 2 corresponds to the classical Lasso algorithm.
- *Direct multi-step predictor identified using the estimated derivative values (PK_ED).* The predictor PK_ED is of the form (25). The basis functions are the same as those used in (25). The derivative values estimated by Algorithm 1 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 $q = 2$, $r = 1$, $\lambda_0 = 1$, $\lambda_i = 200$, $i > 0$, $\Lambda = 50$, and $\rho = 0.4$.

The identified models were tested on the validation set in the task of τ-step ahead prediction, with $\tau \in \{3, 5, 7\}$. The τ-step prediction of models P1_NOD, P1_D and P1_ED was computed by iterating τ times equation (23). The τ-step prediction of models PK_NOD and PK_ED was computed directly using equation (25).

Table 2. Chua circuit. Validation set; $\tau \in \{3, 5, 7\}$. RRMSE prediction errors.

Predictor	RRMSE_3	RRMSE_5	RRMSE_7
P1_NOD	0.035	0.058	0.082
P1_D	3.2e-3	7.3e-3	0.013
P1_ED	2.0e-3	5.7e-3	0.012
PK_NOD	0.034	0.056	0.079
PK_ED	3.5e-4	6.1e-4	1.2e-3

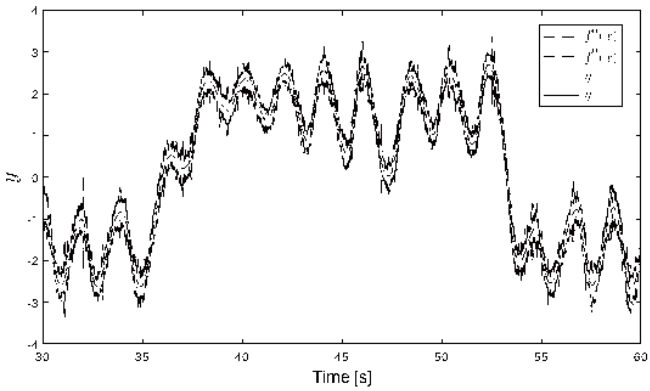

Figure 3. Chua circuit. Validation set (a portion); std = 0.05. 3-step prediction of model PK_ED and related uncertainty bounds.

The results of these tests are summarized in Table 2, where the relative root mean square prediction error RRMSE_τ is reported, for $\tau \in \{3, 5, 7\}$. This performance index is defined as the root mean square (RMS) of the prediction error divided by the RMS of the signal. Figure 3 shows the true system output, the 3-step prediction of the model PK_ED (in the case where std = 0.05) and the related uncertainty bounds, for a portion of the validation set. 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 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. The fact that P1_ED shows a slightly better performance than P1_D appears to be fortuitous. Indeed, we repeated several times the identification/validation procedure, using different noise realizations and the performance of P1_ED and P1_D resulted to be very similar in average. A third observation (important in general but less important than the other two in the context considered in this paper) is that the direct τ-step predictors are in general more accurate than the iterated 1-step predictors.

9.3. Example: control of the inverted pendulum

A pendulum described by the following state equations is considered:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{K}{J}\sin(x_1) - \frac{\beta}{J}x_2 + \frac{1}{J}u \\ y &= x_1 + \xi^c \end{aligned} \tag{26}$$

where the states x_1 and x_2 are the angular position and velocity, respectively, u is the applied torque input, y is the system output, and ξ^c is a disturbance. J is the pendulum's moment of inertia, β is a friction coefficient, and $K = gml$, where g is the gravity acceleration, m is the pendulum's mass and l is its length.

The system (26), discretized via the forward Euler method, can be written in the input-output regression form

$$y_k = a_1 y_{k-1} + a_2 y_{k-2} + a_3 \sin(y_{k-2}) + b_2 u_{k-2} + \xi_k \tag{27}$$

where ξ_k is a noise accounting for the disturbance ξ^c in (26) and b_i, a_i are suitable coefficients, defined from the physical parameters. Equivalently, this equation can be written in the form (1), with $x_k = (y_k, y_{k-1}, u_{k-1})$.

The pendulum equations (26) have been implemented in Simulink, adopting the following parameter values: $J = 0.64$ kg·m^2, $K = 7.848$ kg·$\frac{\text{m}^2}{\text{s}^2}$, $\beta = 0.2$ N·m$\frac{\text{s}}{\text{rad}}$. A command input signal was built in order to make the pendulum work in an "inverted condition", with large movements performed around its unstable equilibrium point. The disturbance ξ^c was simulated as a normally distributed random signal truncated at 3-std, with zero mean and a noise-to-signal standard deviation ratio of 0.05. A simulation of duration 600 s was performed and a set of data of the form (2) was collected, with sampling time $T_s = 0.02$ s, corresponding to an experiment length $L = 30000$. A dataset representative of the system dynamics in the range $180° \pm 70°$, called the identification dataset, was obtained. The corresponding input and output signals are shown in Figure 4.

The following models were identified from the identification dataset.

- *Model identified not using any derivative information (P_NOD).* The model P_NOD is given by

$$\begin{aligned} y_{k+1} &= \hat{f}\left(x_k\right) \\ x_k &= (y_k, y_{k-1}, u_{k-1}) \end{aligned} \tag{28}$$

  where $\hat{f}$ is of the form (5). The basis function set is composed of monomials of degree 3 in the variables y_k, y_{k-1}, u_{k-1}. This set consists of $N = 20$ 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 $\Lambda = 0$. Note that, with these parameter

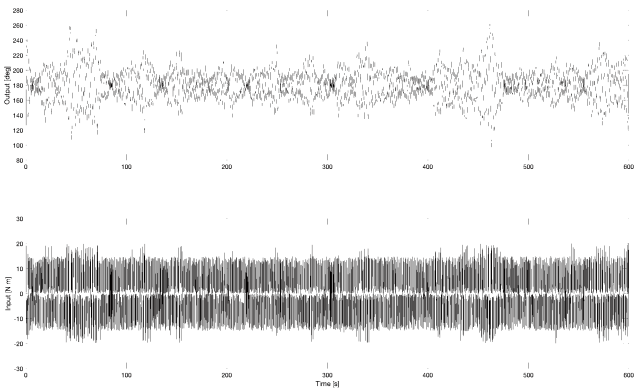

Figure 4. Pendulum. Identification dataset: input and output signals.

values, Method 2 corresponds to the classical Lasso algorithm.

- *Model identified using the true derivative values (P_D).* The model P_D is of the form (28). The basis functions are the same as those used in (28). The true derivative values computed from (27) 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 $q = 2$, $r = 1$, $\lambda_0 = 1$, $\lambda_i = 2$, $i > 0$, and $\Lambda = 0$.
- *Model identified using the estimated derivative values (P_ED).* The model P_ED is of the form (28). The basis functions are the same as those used in (28). The derivative values estimated by Algorithm 1 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 $q = 2$, $r = 1$, $\lambda_0 = 1$, $\lambda_i = 10$, $i > 0$, $\Lambda = 0$, and $\rho = 0.5$.

The choice $\Lambda = 0$ for all models was made on purpose, in order to not confuse the regularization effect given by minimizing the coefficient vector norm with the effect coming from the use of the derivative information. The choice of the parameters λ_i was carried out by means of preliminary tests on a validation set different from the identification set, in the task of τ-step ahead prediction with $\tau = 5$.

From the identified models, three NMPC controllers were designed, named C_NOD, C_D and C_ED, respectively. C_NOD is the controller based on the prediction model P_NOD, C_D is the controller based on P_D and C_ED is the controller based on P_ED.

Then, each controller was applied to the true plant (26), and tuned through closed-loop simulations, using a filtered staircase reference signal, with random values in the range $180° \pm 70°$. On the basis of the simulations, the following parameters where chosen for all the controllers: sampling time $T_s = 0.02$ s, prediction horizon $T_p = 5T_s$ s, control horizon $T_c = T_s$ s, and weights $Q = 1$ and $R = 10^{-6}$. Note that the controllers were tuned independently from each other, in order to obtain the best possible performance for each controller. Such an "independent" tuning led to the choice of the same parameters for all the controllers.

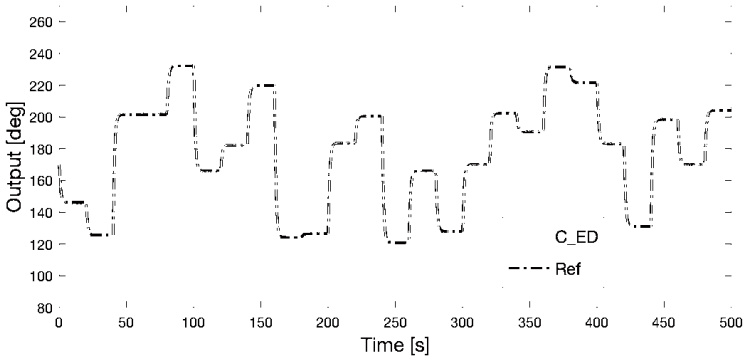

Figure 5. Pendulum. C_ED controller. Reference versus actual closed-loop output.

The controllers were tested on the true plant (26), through closed-loop simulations. The following tests were carried out:

- Staircase 1: filtered staircase reference signal, with random values in the range $180° \pm 60°$, starting from the initial angle 180° with null speed.
- Staircase 2: filtered staircase reference signal, with random values in the range $180° \pm 70°$, starting from the initial angle 180° with null speed.
- Steps: non-filtered step reference signals, with random values in the range $180° \pm 40°$, starting from the initial angle 180° with null speed.

Note that all the reference signals used for these tests are different from those used for model identification and controller tuning. These signals have been chosen to make the pendulum work in an "inverted modality", with large movements around its unstable equilibrium point. These kinds of maneuvers are indeed more challenging than maneuvers operated near the stable equilibrium point. Figure 5 shows the comparison between the reference signal and the output of the closed-loop system with the C_ED controller in the Staircase 1 test. In the Staircase 2 test and in step tests with reference values around $180° \pm 40°$ or larger, the C_NOD controller yielded a divergent behavior, while the other two controllers worked correctly. Table 3 shows the results obtained in the staircase tests and in some step tests where the C_NOD controller did not lead to a divergent behavior.

From these results, it can be observed that the C_D controller, which is based on the exact information about the system function derivatives, gives the best performance, both in steady-state and transient conditions. The C_NOD controller, not using any information about the derivatives, shows the worst performance, with divergent behaviors in the case of large and/or non-smooth reference signals. The C_ED controller, based on an approximated information about the derivatives, provides an intermediate performance, in any case significantly better than the one given by the C_NOD controller.

Table 3. Pendulum. Control performance indices. OS: overshoot [%]; RT: rise time [s]; ST: settling time 2% [s]; SSE: steady-state tracking error [deg]; RMSE: root mean square tracking error [deg].

	index	C_NOD	C_D	C_ED
Staircase 1	RMSE	1.82	0.59	1.33
Staircase 2	RMSE	∞	0.68	1.46
Step 1	OS	17	5.8	8.3
	SSE	0.362	0.002	0.22
	RT	0.061	0.064	0.065
	ST	0.748	0.31	0.407
Step 2	OS	36.8	3.9	12.6
	SSE	0.593	0.031	0.369
	RT	0.053	0.067	0.061
	ST	0.74	0.205	0.386
Step 3	OS	23	14.6	19.1
	SSE	0.272	0.076	0.184
	RT	0.1	0.104	0.104
	ST	0.621	0.319	0.448

10. Conclusions

An approach for the identification of a function together with its derivatives has been proposed in this paper. Within this approach, an optimality analysis has been developed, guaranteed uncertainty bounds have been derived, and a technique for estimating the derivative values from the input-output data has been presented. The approach has been tested on simulated examples concerned with identification of a univariate function, multi-step prediction of the Chua chaotic circuit and nonlinear model predictive control of a pendulum. In these examples, the models identified using the proposed methods resulted to be significantly more accurate than models obtained using a standard identification technique, thus demonstrating the potential of the proposed identification approach. The application of the proposed identification method in the context of nonlinear predictive control appears to be particularly promising. Besides NMPC, future research activities will be dedicated to a deeper investigation on the use of direct multi-step predictors and to the study/developement of alternative algorithms for the estimation of the derivative samples.

References

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

Chen, J., & Gu, G. (2000). *Control-oriented system identification: An H_∞ approach*. New York: John Wiley & Sons.

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

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

Donoho, D., Elad, M., & Temlyakov, V. (2006, jan.). Stable recovery of sparse overcomplete

representations in the presence of noise. *IEEE Transactions on Information Theory*, *52*(1), 6 - 18.

Findeisen, R., Allgower, F., & Biegel, L. (2007). Assessment and future directions of nonlinear model predictive control. In *Lecture notes in control and information sciences.* Springer.

Freeman, A., & Kokotovic, V. (1996). *Robust nonlinear control design.* Boston: Birkhï¿œuser.

Fuchs, J. (2005, oct.). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, *51*(10), 3601 -3608.

Goodwin, G., & Ninness, B. (1992). Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Transactions on Automatic Control*, *37*(7), 913-928.

Goodwin, G., Yuz, J., Aguero, J., & 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., & Pannek, J. (2011). Nonlinear model predictive control - theory and algorithms. In *Communications and control engineering.* Springer.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks.*

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

Magni, L., Raimondo, D., & Allgower, F. (2009). Nonlinear model predictive control - towards new challenging applications. In *Lecture notes in control and information sciences.* Springer.

Mai-Duy, N., & Tran-Cong, T. (2003). Approximation of function and its derivatives using radial basis function networks. *Applied Mathematical Modelling.*

Manzano, J., Limon, D., de la Peñ, D. M., & 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., & Walter, E. (1996). *Bounding approaches to system identification.* Plenum Press.

Milanese, M., & Novara, C. (2004). Set membership identification of nonlinear systems. *Automatica*, *40/6*, 957-975.

Milanese, M., & Novara, C. (2005). Set membership prediction of nonlinear time series. *IEEE Transactions on Automatic Control*, *50*(11), 1655-1669.

Milanese, M., & Novara, C. (2011). Unified set membership theory for identification, prediction and filtering of nonlinear systems. *Automatica*, *47*(10), 2141-2151.

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

Milanese, M., & Vicino, A. (1993). Information-based complexity and nonparametric worst-case system identification. *Journal of Complexity*, *9*, 427-446.

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., & Milanese, M. (2016). Data-driven design of two degree-of-freedom nonlinear controllers: the D2-IBC approach. *Automatica*, *72*, 19-27.

Novara, C., & Milanese, M. (2019). Control of mimo nonlinear systems: A data-driven model inversion approach. *Automatica*, *101*, 417 - 430. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0005109818306332`

Novara, C., Vincent, T., Hsu, K., Milanese, M., & Poolla, K. (2011). Parametric identification of structured nonlinear systems. *Automatica*, *47*(4), 711 - 721.

Piga, D., Forgione, M., Formentin, S., & Bemporad, A. (2019). Performance-oriented model learning for data-driven mpc design. *IEEE Control Systems Letters*, *3*(3), 577-582.

Pillonetto, G., & De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, *46*(1), 81-93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., & Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, *50*(3), 657-682.

Popper, K. R. (1969). *Conjectures and refutations: The growth of scientific knowledge.* London: Rontedge and Kegan Paul.

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

Qu, Z. (1998). *Robust control of nonlinear uncertain systems.* Wiley series in nonlinear science.

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

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

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

Sznaier, M., Wenjing, M., Camps, O., & 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., & Woźniakowski, H. (1988). *Information-based complexity.* Academic Press, Inc.

Tropp, J. (2006, mar.). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, *52*(3), 1030 -1051.

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

Appendix: Theorem proofs

Proof of Theorems 1 and 2. If the optimization problem (8)-(9) is feasible, then an approximation $\hat{f}$ of the form (5) exists, such that inequalities (9) are satisfied. These inequalities are equivalent to the following ones: $||\tilde{z}_i - \hat{f}^{(i)}(\tilde{x})||_q \leq \mu_i$, $i = 0, \ldots, n_x$. Moreover, $\hat{f} \in \mathcal{S}_{1p}(X)$ by definition. It follows that $\hat{f} \in \mathrm{FFS}_{\mathcal{S}}$, which implies $\mathrm{FFS}_{\mathcal{S}} \neq \emptyset$. This proves Theorem **1.**

As shown in (Milanese & Vicino, 1993), equation (11), if $\hat{f} \in \mathrm{FFS}_{\mathcal{S}}$, then $\hat{f}$ is $\mathrm{FFS}_{\mathcal{S}}$-almost-optimal. This proves Theorem 2. □

Proof of Theorems 3 and 4. The proof of Theorem 1 shows that, if the optimization problem (8)-(9) is feasible, then an approximation $\hat{f}$ of the form (5) exists, and $\hat{f} \in \mathrm{FFS}_{\mathcal{S}}$. Consider now the function $f = \hat{f} + \Delta$, with $\Delta = 0$. Obviously, $f = \hat{f} \in \mathrm{FFS}_{\mathcal{S}}$ and $f^{(i)} - \hat{f}^{(i)} = \Delta = 0 \in \mathcal{L}(\gamma_i, X)$, for any $\gamma_i \geq 0$. From Definitions 2 and 7, it follows that $f = \hat{f} \in \mathrm{FFS}_{\mathcal{L}}$, which implies $\mathrm{FFS}_{\mathcal{L}} \neq \emptyset$. This proves Theorem 3.

As shown in (Milanese & Vicino, 1993), equation (11), if $\hat{f} \in \mathrm{FFS}_{\mathcal{L}}$, then $\hat{f}$ is $\mathrm{FFS}_{\mathcal{L}}$-almost-optimal. This proves Theorem 4. □

Proof of Theorem 5. The proof for the case $i = 0$ comes from Theorem 3 in (Novara, 2016). This theorem shows that the following bounds hold for every $x \in X$:

$$f_o(x) \leq \overline{f}_0(x) \equiv \hat{f}(x) + \overline{\Delta}_0(x)$$
$$f_o(x) \geq \underline{f}_0(x) \equiv \hat{f}(x) + \underline{\Delta}_0(x).$$

In the case $i > 0$, under the assumption (5.2), we can follow the same argumentations of the proof of Theorem 3 in (Novara, 2016). In this way, we obtain that the following bounds hold for every $x \in X$:

$$\begin{aligned} f_o^{(i)}(x) &\leq \hat{f}^{(i)}(x) + \overline{\Delta}_i(x) \\ f_o^{(i)}(x) &\geq \hat{f}^{(i)}(x) + \underline{\Delta}_i(x). \end{aligned} \tag{29}$$

Moreover, we know that $\Delta^{(0)}$ is Lipschitz continuous with constant γ_0. This implies that

$$\left| f_o^{(i)} - \hat{f}^{(i)}(x) \right| \leq \gamma_0 \equiv \bar{\gamma}, \ i = 1, \ldots, n_x. \tag{30}$$

The bounds (13) for $i > 0$ are obtained from (29) and (30). Equations (14) follow from Theorem 2 in (Milanese & Novara, 2004). □

Proof of Theorem 6. Let us consider the Taylor expansion of f_o around a point $\tilde{x}_k$:

$$f_o(x) = f_o(\tilde{x}_k) + (x - \tilde{x}_k)^\top \nabla f_o(\tilde{x}_k) + R(x - \tilde{x}_k)$$

where $\nabla f_o = (f_o^{(1)}, \ldots, f_o^{(n_x)})$ is the gradient of f_o and $R(\cdot)$ is a reminder. This expression, evaluated at a point $\tilde{x}_j$, with $j \in \Upsilon_{\rho k}$, becomes

$$f_o(\tilde{x}_j) = f_o(\tilde{x}_k) + (\tilde{x}_j - \tilde{x}_k)^\top \nabla f_o(\tilde{x}_k) + R(\tilde{x}_j - \tilde{x}_k).$$

From (3), this can be written as

$$\tilde{z}_j - \tilde{z}_k = (\tilde{x}_j - \tilde{x}_k)^\top \nabla f_o(\tilde{x}_k) + R(\tilde{x}_j - \tilde{x}_k) + d_j - d_k.$$

For $j = j_1, \ldots, j_M$, we obtain the following equation in matrix form:

$$\tilde{z}_{\rho k} = \Phi_{\rho k} \nabla f_o(\tilde{x}_k) + \Xi_k + D_k$$

where $\Xi_k \doteq (R(\tilde{x}_{j_1} - \tilde{x}_k), \ldots, R(\tilde{x}_{j_M} - \tilde{x}_k))$ and $D_k \doteq (d_{j_1} - d_k, \ldots, d_{j_M} - d_k)$. It follows that

$$\nabla f_o(\tilde{x}_k) = \Phi_{\rho k}^\dagger \tilde{z}_{\rho k} - \Phi_{\rho k}^\dagger (\Xi_k + D_k)$$

where $\Phi_{\rho k}^\dagger \doteq (\Phi_{\rho k}^\top \Phi_{\rho k})^{-1} \Phi_{\rho k}^\top$. The inverse $(\Phi_{\rho k}^\top \Phi_{\rho k})^{-1}$ exists and is finite since

$\frac{1}{M}\Phi_{\rho k}^\top \Phi_{\rho k} \succ 0$, $\forall M \geq M_0$, by assumption. This matrix inequality also implies that the solution of the optimization problem (17) is given by $g_k = \Phi_{\rho k}^\dagger \tilde{z}_{\rho k}$. The vector g_k is an estimate of the gradient $\nabla f_o(\tilde{x}_k)$. The resulting estimation error $\nabla f_o(\tilde{x}_k) - g_k$ is bounded as

$$\begin{aligned}
&\|\nabla f_o(\tilde{x}_k) - g_k\|_q = \|\Phi_{\rho k}^\dagger (\Xi_k + D_k)\|_q \\
&\leq \|\Phi_{\rho k}^\dagger\|_q \|\Xi_k + D_k\|_q \leq \|\Phi_{\rho k}^\dagger\|_q \left(\|\Xi_k\|_q + 2\mu_0\right).
\end{aligned}$$

Being $f_o^{(i)}$ Lipschitz continuous by assumption, each element of Ξ_k is bounded as

$$|R(\tilde{x}_j - \tilde{x}_k)| \leq \gamma_R \|\tilde{x}_j - \tilde{x}_k\|_q \leq \rho\gamma_R, \; \forall \tilde{x}_j \in X$$

for some $\gamma_R \geq 0$, $\gamma_R < \infty$. It follows that, for $\forall M \geq M_0$,

$$\|\Xi_k\|_q \leq \begin{cases} \rho\sqrt{M}\gamma_R, & q = 2 \\ \rho\gamma_R, & q = \infty. \end{cases} \tag{31}$$

Hence,

$$\begin{aligned}
&\|\nabla f_o(\tilde{x}_k) - g_k\|_q \leq \|\Phi_{\rho k}^\dagger\|_q \|\Xi_k\|_q + \|\Phi_{\rho k}^\dagger\|_q + 2\mu_0 \\
&\leq \|\Phi_{\rho k}^\dagger\|_q \rho\sqrt{M}\gamma_R + \|\Phi_{\rho k}^\dagger\|_q + 2\mu_0 \; (q = 2) \\
\text{or } &\leq \|\Phi_{\rho k}^\dagger\|_q \rho\gamma_R + \|\Phi_{\rho k}^\dagger\|_q + 2\mu_0 \; (q = \infty).
\end{aligned}$$

The statement is proven choosing $\rho = \epsilon/(\|\Phi_{\rho k}^\dagger\|_q \sqrt{M}\gamma_R)$ $(q = 2)$ or $\rho = \epsilon/(\|\Phi_{\rho k}^\dagger\|_q \gamma_R)$ $(q = \infty)$. □

Proof of Theorem 7. Let us denote the function gradient as $g_o \doteq \nabla f_o(\tilde{x}_k)$ and, for a certain gradient estimate g, the estimation error as $\delta g \doteq g_o - g$. The objective function of the optimization problem (17) is

$$J(g) \doteq \frac{1}{M}\|\tilde{z}_{\rho k} - \Phi_{\rho k} g\|_2^2.$$

This function can be written as

$$\begin{aligned}
J(g) &= \frac{1}{M}(\tilde{z}_{\rho k} - \Phi_{\rho k} g)^\top (\tilde{z}_{\rho k} - \Phi_{\rho k} g) \\
&= \frac{1}{M}\left(\tilde{z}_{\rho k} - \Phi_{\rho k} g_o + \Phi_{\rho k}\delta g\right)^\top \left(\tilde{z}_{\rho k} - \Phi_{\rho k} g_o + \Phi_{\rho k}\delta g\right) \\
&= \frac{1}{M}\left(\Xi_k + D_k + \Phi_{\rho k}\delta g\right)^\top \left(\Xi_k + D_k + \Phi_{\rho k}\delta g\right) \\
&= \frac{1}{M}\Xi_k^\top \Xi_k + \frac{1}{M}D_k^\top D_k + \frac{1}{M}\delta g^\top \Phi_{\rho k}^\top \Phi_{\rho k}\delta g \\
&+ \frac{2}{M}D_k^\top \Xi_k + \frac{2}{M}\Xi_k^\top \Phi_{\rho k}\delta g + \frac{2}{M}D_k^\top \Phi_{\rho k}\delta g.
\end{aligned}$$

From (31) and the noise bounds $\|d_i\|_q \leq \mu_i$, a sufficiently large M_0 exists such that

$$\begin{array}{l} \frac{1}{M}\Xi_k^\top \Xi_k \leq \gamma_R^2 \rho^2, \ \forall M \geq M_0 \\ \frac{1}{M}\left|D_k^\top \Xi_k\right| \leq 2\breve{\mu}_{0,i}\gamma_R \rho, \ \forall M \geq M_0. \end{array}$$

From (19) and (20), for every $\epsilon > 0$, a sufficiently large M_0 exists such that

$$\begin{array}{l} \left|\dfrac{1}{M}D_k^\top D_k - \sigma_D^2\right| \leq \epsilon, \ \forall M \geq M_0 \\ \left|\dfrac{1}{M}D_k^\top \Phi_{\rho k}\delta g\right| \leq \|\delta g\|_2 \epsilon, \ \forall M \geq M_0. \end{array}$$

Moreover,

$$\frac{1}{M}\left|\Xi_k^\top \Phi_{\rho k}\delta g\right| \leq \frac{1}{\sqrt{M}}\|\Phi_{\rho k}\|_2 \|\delta g\|_2 \gamma_R \rho.$$

The quantity $\|\Phi_{\rho k}\|_2/\sqrt{M}$ is bounded as

$$\begin{array}{l} \frac{1}{\sqrt{M}}\|\Phi_{\rho k}\|_2 \leq \frac{1}{\sqrt{M}}\left(\sum_{j=1}^{M}\sum_{i=1}^{n_x}(\Phi_{\rho k})_{ji}^2\right)^{1/2} \\ \leq \frac{1}{\sqrt{M}}\left(n_x M \max_{i,j}(\Phi_{\rho k})_{ji}^2\right)^{1/2} = \sqrt{n_x}\max_{i,j}\left|(\Phi_{\rho k})_{ji}\right| \end{array}$$

where the first inequality is a standard result in the literature and $(\Phi_{\rho k})_{ji}$ are the entries of $\Phi_{\rho k}$. Note that $\max_{i,j}\left|(\Phi_{\rho k})_{ji}\right|$ is bounded, since the measurements $\tilde{x}_j$ are assumed to be in a compact set. The quantity $\|\delta g\|_2$ is bounded on any compact set G containing g_o: for all $g \in G$, $\|\delta g\|_2 \leq \bar{G}$, for some $\bar{G} > 0$, $\bar{G} < \infty$.

From all the above inequalities, we have that

$$\begin{aligned} |J(g) - J_o(g)| \leq \gamma_R^2\rho^2 + 4\breve{\mu}_{0,i}\gamma_R\rho + \epsilon + 2\bar{G}\epsilon \\ + 2\sqrt{n_x}\max_{i,j}\left|(\Phi_{\rho k})_{ji}\right|\bar{G}\gamma_R\rho \end{aligned}$$

where

$$J_o(g) \doteq \frac{1}{M}\delta g^\top \Phi_{\rho k}^\top \Phi_{\rho k}\delta g + \sigma_D^2.$$

It follows that, as $\rho \to 0$ and $M \to \infty$, $J(g)$ converges to $J_o(g)$.

This convergence is uniform on any compact set G containing g_o. It follows that the minimizers of $J(g)$ converge to the minimizers of $J_o(g)$, see (Ljung, 1999). The condition $\frac{1}{M}\Phi_{\rho k}^\top \Phi_{\rho k} \succ 0$ ensures that $J_o(g)$ has a unique minimizer, given by $g_o \doteq \nabla f_o(\tilde{x}_k)$. The claim follows. □