2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Sparse identification of nonlinear functions and parametric Set Membership optimality analysis

Carlo Novara

Abstract— **Identifying a sparse approximation of a function from a set of data can be useful to solve relevant problems in the automatic control field. However, finding a sparsest approximation is in general an NP-hard problem. The common approach is to use relaxed or greedy algorithms that, under certain conditions, can provide sparsest solutions. In this paper, a combined ℓ_1-relaxed-greedy algorithm is proposed and a condition is given, under which the approximation derived by the algorithm is a sparsest one. Differently from other conditions available in the literature, the one provided here can be easily verified for any choice of the basis functions. A Set Membership analysis is also carried out assuming that the function to approximate is a linear combination of unknown basis functions belonging to a known set of functions. It is shown that the algorithm is able to exactly select the basis functions which define the unknown function and to provide an optimal estimate of their coefficients. It must be remarked that exact basis function selection is performed for a finite number of data, whereas in standard system identification, a similar result can only be obtained for an infinite number of data. A simulation example, related to the identification of vehicle lateral dynamics, is finally presented.**

I. Introduction

Sparse approximation consists in approximating a function using a “few” basis functions properly selected within a “large” set. More precisely, a sparse approximation is a linear combination of “many” basis functions, but the vector of linear combination coefficients is “sparse”, i.e. it has only a “few” non-zero elements. Deriving a sparse approximation of an unknown function from a set of its values (possibly corrupted by noise) is here called sparse identification.

Sparsification methods are relevant in many applications: compressive sensing [1], [2], [3], bioinformatics [4], computer vision [5], signal processing [6], [7], [8], source separation [9], denoising [10], linear regression [11], and regularization [12]. Analogies between sparse approximation and support vector machines have been shown in [13]. Recently, sparsification methods have been introduced in the automatic control field [14], [15], [16], with promising results. In this field, applications might include regularization, basis function selection (see Sections V and VI), regressor selection, input selection, nonlinear internal model control, nonlinear feed-forward control, direct inverse control, and approximation of predictive controllers for fast online implementation.

The sparsity of a vector is typically measured by the ℓ_0 quasi-norm, defined as the number of its non-zero elements. Sparse identification can be performed by looking for a coefficient vector of the basis function linear combination with a “small” ℓ_0 quasi-norm that yields a “small” prediction error evaluated on the measured data. However, the ℓ_0 quasi-norm is a non-convex function and its minimization is in general an NP-hard problem. Two main approaches are commonly adopted to deal with this issue: convex relaxation and greedy algorithms, [17], [18], [19], [20]. In convex relaxation, a suitable convex function is minimized instead of the ℓ_0 quasi-norm. In greedy algorithms, the sparse solution is obtained iteratively, [17]. A very interesting feature of these approaches is that, under certain conditions, they provide sparsest solutions, i.e. solutions which also minimize the ℓ_0 quasi-norm [17], [18], [19], [21]. Although such conditions give an important theoretical motivation for using these relaxed/greedy approaches, their actual verification is often hard from a computational point of view. A remarkable contribution on this topic was provided in [21], [18], [20]. In [21], the conditions for a vector to be the sparsest solution can be easily verified when the basis functions are orthonormal or the union of incoherent orthonormal bases. In [18] and [20], the conditions are of easy verification, but require that the basis functions have “small” mutual coherence.

In this paper, a combined relaxed-greedy algorithm is proposed for sparse identification and a condition is provided, under which the solution derived by the algorithm is sparsest. Such a condition is easily verifiable for any choice of the basis functions. A bound on the number of non-zero elements of the algorithm solution which may be in excess with respect to the sparsest solution is also derived.

A Set Membership optimality analysis is then performed in order to assess the accuracy of the approximation obtained by the relaxed-greedy algorithm. The noise affecting the data is assumed bounded in norm and, as common in system identification, [22], [23], [24], the unknown function is assumed to be a linear combination of basis functions whose coefficients have to be estimated. It is supposed that the basis functions are not known but belong to a known set of functions. It is shown that the algorithm is able to select exactly the basis functions defining the unknown function and to provide an optimal estimate of their coefficients. It must be remarked that the exact basis function selection is performed for a finite number of data, whereas in standard system identification, a similar result can be obtained only for an infinite number of data [22], [23], [24].

Finally, a simulation example is shown, related to the identification of vehicle lateral dynamics.

C. Novara is with Dip. di Automatica e Informatica, Politecnico di Torino, Italy, carlo.novara@polito.it

978-1-4577-0079-8/11/$26.00 ©2011 AACC

II. Notation and basic notions

A column vector is indicated by $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^{n\times 1}$, a row vector by $a^T = [a_1, a_2, \ldots, a_n] \in \mathbb{R}^{1\times n}$. For a matrix/vector $A \in \mathbb{R}^{K\times n}$, $K \in \{1, 2, \ldots\}$, and a set of indices $\lambda = \{i_1, i_2, \ldots, i_m\} \subset \{1, 2, \ldots, n\}$, let us introduce the notation

$$A_\lambda \doteq [A_{i_1}, A_{i_2}, \ldots, A_{im}]$$

where A_j are the columns/elements of A.

The ℓ_q norm of a vector a is defined as

$$\begin{aligned} \|a\|_q &\doteq \left(\textstyle\sum_{i=1}^n |a_i|^q\right)^{\frac{1}{q}}, \ q \in [1, \infty), \\ \|a\|_\infty &\doteq \max_{i=1,..,n} |a_i| . \end{aligned}$$

The ℓ_0 quasi-norm of a vector $a \in \mathbb{R}^n$ is defined as the number of its elements which are not null:

$$\|a\|_0 \doteq \operatorname{card}(\operatorname{supp}(a)) \tag{1}$$

where $\operatorname{card}(\cdot)$ is the set cardinality, and $\operatorname{supp}(a)$ is the support of a, defined as the set of indices at which a is not null:

$$\operatorname{supp}(a) \doteq \{i \in \{1, 2, \ldots, n\} : a_i \neq 0\} .$$

The ℓ_0 quasi-norm is commonly used to measure the *sparsity* of a vector: the smaller is the ℓ_0 quasi-norm, the sparser is the vector. The complement of $\operatorname{supp}(a)$, i.e. the set of indices at which a is null, is denoted by

$$\begin{aligned} \overline{\operatorname{supp}}(a) &\doteq \{i \in \{1, 2, \ldots, n\} : a_i = 0\} \\ &= \{1, 2, \ldots, n\} \setminus \operatorname{supp}(a) . \end{aligned}$$

The L_p norm of a function $f : X \to Y$, where $X \subseteq \mathbb{R}^{n_x}$ and $Y \subseteq \mathbb{R}$, is defined as

$$\begin{aligned} \|f\|_p &\doteq \left[\textstyle\int_X |f(x)|^p \, dx\right]^{\frac{1}{p}}, \ p \in [1, \infty), \\ \|f\|_\infty &\doteq \operatorname{ess\,sup}_{x\in X} |f(x)| . \end{aligned}$$

Consider a generic optimization problem

$$\begin{aligned} &a = \arg\min_a J(a) \\ &\text{subject to} \quad g(a) \leq 0. \end{aligned}$$

If this problem admits a set of solutions, then a indicates one of these solutions. Otherwise, a is the unique solution.

III. Problem formulation

Consider a nonlinear function f_0 defined by

$$y = f_0(x) \tag{2}$$

where $x \in X \subset \mathbb{R}^{n_x}$, $y \in Y \subset \mathbb{R}$. Suppose that f_0 is not known but a set of noise-corrupted data $D = \{\widetilde{x}_k, \widetilde{y}_k\}_{k=1}^L$ is available, described by

$$\widetilde{y}_k = f_0(\widetilde{x}_k) + d_k, \quad k = 1, 2, \ldots, L \tag{3}$$

where d_k is noise. Define the following parameterized function:

$$f_a(x) = \sum_{i=1}^n a_i \phi_i(x) = \phi(x)\, a \tag{4}$$

where $\phi(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_n(x)]$, $\phi_i : X \to Y$ are known basis functions, and $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ is an unknown coefficient vector.

Problem 1: From the data set D, identify a coefficient vector a such that
(i) a is "sparse",
(ii) the identification error

$$e(f_a) \doteq \|f_0 - f_a\|_p \tag{5}$$

is "small". ■

IV. Sparse identification of nonlinear functions

A possible solution to the sparse identification problem (Problem 1) may be obtained by looking for the sparsest coefficient vector which guarantees a given maximum error ε between the measured output $\widetilde{y}$ and the predicted output $f_a(\widetilde{x}) = \phi(\widetilde{x})\, a$. According to (1), minimizing the ℓ_0 quasi-norm of a vector corresponds to minimizing the number of its non-zero elements, i.e. to maximizing its sparsity. Thus, a solution to Problem 1 could be found by solving the following optimization problem:

$$\begin{aligned} &a^0 = \arg\min_{a\in\mathbb{R}^n} \|a\|_0 \\ &\text{subject to} \quad \|\widetilde{y} - \Phi a\|_2 \leq \varepsilon \end{aligned} \tag{6}$$

where

$$\begin{aligned} \widetilde{y} &\doteq (\widetilde{y}_1, \ldots, \widetilde{y}_L) \\ \Phi &\doteq \begin{bmatrix} \phi_1(\widetilde{x}_1) & \cdots & \phi_n(\widetilde{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\widetilde{x}_L) & \cdots & \phi_n(\widetilde{x}_L) \end{bmatrix} \\ &= \begin{bmatrix} \phi_1(\widetilde{x}) & \cdots & \phi_n(\widetilde{x}) \end{bmatrix} . \end{aligned}$$

However, the ℓ_0 quasi-norm is a non-convex function and its minimization is in general an NP-hard problem. Two main approaches are commonly adopted to deal with this issue: convex relaxation and greedy algorithms, [17], [18], [19], [20]. In convex relaxation, an optimization problem similar to (6) is solved, where the ℓ_0 quasi-norm is replaced by a suitable convex function. The ℓ_1 norm is often used, since this norm is the *convex envelope* of the ℓ_0 quasi-norm. In greedy algorithms, a sparse solution is obtained iteratively, by successively individuating the most "important" vector elements, [17]. Under certain conditions, these approaches give sparsest solutions, i.e. solutions which are also solution of (6), [17], [18], [19], [21]. However, the verification of these conditions is in general hard from a computational standpoint, and can actually be performed only for particular types of basis functions.

In the remaining of this section, a combined ℓ_1-relaxed-greedy algorithm is proposed for solving the sparse identification problem 1. A theorem is presented, giving a condition easily verifiable for any basis functions, under which the solution derived by the algorithm is the sparsest one.

Without loss of generality, assume that the columns of Φ are normalized: $\|\phi_i(\widetilde{x})\|_2 = 1$, $i = 1, 2, \ldots, n$. For a given vector $w \in \mathbb{R}^L$, define the following norm:

$$|w|_K \doteq \sqrt{\sum_{i \in I_K} \left(w^T \phi_i\left(\widetilde{x}\right)\right)^2}$$

where I_K is the set of the K largest inner products $w^T \phi_i\left(\widetilde{x}\right)$. Define also the following quantities:

$$\begin{array}{l} \delta\left(a\right) \doteq \widetilde{y} - \Phi a \\ \xi\left(a\right) \doteq \frac{|\delta(a)|_1 + |\delta(a)|_{2m}}{\underline{\sigma}^2(\Phi)} \\ r\left(a\right) \doteq \left\{i_1, \ldots, i_j : |a_{i_1}| \geq \ldots \geq \left|a_{i_j}\right| \geq \xi\left(a\right)\right\} \end{array} \tag{7}$$

where $\underline{\sigma}\left(\Phi\right)$ is the minimum non-zero singular value of Φ and $m\left(a\right) \doteq \|a\|_0$.

Algorithm 1:

1) Solve the optimization problem

$$\begin{array}{l} a^1 = \arg\min\limits_{a \in \mathbb{R}^n} \|a\|_1 \\ \text{subject to} \quad \|\widetilde{y} - \Phi a\|_2 \leq \varepsilon \end{array} \tag{8}$$

2) Compute the coefficient vector a^* as follows:

```
for $k = 1 : n - m\left(a^1\right)$

    $c^k = \arg\min_{a \in \mathbb{R}^n} \|\widetilde{y} - \Phi a\|_2$
    subject to    $a_i = 0, \ \forall i \in r_\lambda\left(a^1\right)$
                  $\lambda = \left\{m\left(a^1\right) + k, \ldots, n\right\}$

    if $\left\|\delta\left(c^k\right)\right\|_2 \leq \varepsilon$
        $a^* = c^k$
        break
    end

end
```

■

The above algorithm provides an estimate a^* of a^0, the solution of the non-convex optimization problem (6). The following theorem gives an easily verifiable condition ensuring that a^* has the same support as a^0. The theorem also allows the computation of a bound on the number of non-zero elements of a^* that are in excess with respect to a^0.

Define the following coefficient vector:

$$\begin{array}{ll} c^{ver} = \arg\min\limits_{a \in \mathbb{R}^n} \|a_\varsigma\|_1 & \\ \quad \text{subject to} & \text{sign}\left(a_i^*\right) a_i \geq \eta\left(a^*\right), \ \forall i \in \varsigma \\ & |a_i| < \eta\left(a^*\right), \ \forall i \in \overline{\varsigma} \\ & \|\widetilde{y} - \Phi a\|_2 \leq \varepsilon \end{array} \tag{9}$$

where

$$\begin{array}{rcl} \eta\left(a\right) & \doteq & \min\limits_{i \in \text{supp}(a)} |a_i| \\ \varsigma & \doteq & \text{supp}\left(a^*\right) \\ \overline{\varsigma} & \doteq & \overline{\text{supp}}\left(a^*\right). \end{array} \tag{10}$$

Theorem 1: Let a^* be the coefficient vector derived by Algorithm 1. Let $N_e \doteq \|a^*\|_0 - \left\|a^0\right\|_0$ and $\lambda^s \doteq \{i : |c_i^{ver}| > \xi\left(c^{ver}\right)\}$. Then,

$$N_e \leq \overline{N}_e \doteq \|a^*\|_0 - \text{card}\left(\lambda^s\right). \tag{11}$$

Moreover, if

$$\xi\left(c^{ver}\right) < \eta\left(a^*\right) \tag{12}$$

then $\overline{N}_e = 0$ and

$$\text{supp}\left(a^*\right) = \text{supp}\left(a^0\right). \tag{13}$$

Proof. See [25]. ■

Verification of the condition (12) in Theorem 1 is computationally simple for any matrix Φ. Indeed, from (7) and (10), this verification basically requires to evaluate $|\delta\left(c^{ver}\right)|_1$, $|\delta\left(c^{ver}\right)|$, $\underline{\sigma}^2\left(\Phi\right)$ and to solve the convex optimization problem (9). All these operations can be easily be performed in polynomial time. It must be remarked that condition (12) has been derived from condition (12) of Corollary 1 in [21], whose verification is simple when the basis functions are orthonormal or the union of incoherent orthonormal bases.

V. Parametric Set Membership optimality analysis and exact basis function selection

In Section IV, an ℓ_1-relaxed-greedy algorithm is presented, able to derive a "sparse" approximation of the function f_0, thus allowing the accomplishment of the requirement (i) of Problem 1. In this section, considering a Set Membership framework [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], this approximation is shown to have "small" identification error, thus allowing us to satisfy also the requirement (ii) of Problem 1.

In order to have a bounded identification error, some assumptions have to be made on the noise affecting the data and on the unknown function f_0. In this paper, the noise sequence $d = \left(d_1, d_2, \ldots, d_L\right)$ in (3) is assumed to be bounded as

$$\|d\|_2 \leq \varepsilon. \tag{14}$$

As common in system identification, [22], [23], [24], the function f_0 is assumed to be parameterized as

$$f_0\left(x\right) = \sum_{i=1}^{n} a_i^0 \phi_i\left(x\right).$$

where $\phi_1\left(x\right), \phi_2\left(x\right), \ldots, \phi_n\left(x\right)$ are known basis functions and $a^0 = \left(a_1^0, a_2^0, \ldots, a_n^0\right) \in \mathbb{R}^n$ is a "sparse" unknown parameter vector, solution of (6).

Under these assumptions, the identification Problem 1 reduces to finding an estimate $\widehat{a}$ of a^0 such that
(i) $\text{supp}\left(\widehat{a}\right) = \text{supp}\left(a^0\right)$,
(ii) the parametric error

$$e^{par}\left(\widehat{a}\right) \doteq \left\|a^0 - \widehat{a}\right\|_2$$

is "small".

While Theorem 1 gives a condition under which $\text{supp}\left(\widehat{a}\right) = \text{supp}\left(a^0\right)$, no exact knowledge of $e^{par}\left(\widehat{a}\right)$ is available, being a^0 unknown. However, under the above assumptions, we have that $a^0 \in FPS$, where

$$\begin{aligned} FPS \doteq \{a \in \mathbb{R}^n : \operatorname{supp}(a) = \operatorname{supp}\left(a^0\right), \\ \left\|\widetilde{y} - \Phi a\right\|_2 \leq \varepsilon\}. \end{aligned}$$

The set FPS is called *Feasible Parameter Set* and is the set of all parameter vectors consistent with the prior assumptions and the measured data. A tight bound on $e^{par}(\widehat{a})$ is thus given by the following *worst-case parametric error*:

$$EP(\widehat{a}) \doteq \sup_{a \in FPS} \left\|a - \widehat{a}\right\|_2 .$$

This leads to the notion of *optimal estimate*, defined as an estimate a^c which minimizes the worst-case parametric error:

$$EP(a^c) = \inf_{\widehat{a} \in \mathbb{R}^n} \sup_{a \in FPS} \left\|a - \widehat{a}\right\|_2 . \tag{15}$$

The following result shows that, under the assumption of Theorem 1, the estimate a^* provided by Algorithm 1 is optimal.

Theorem 2: Let a^* be the parameter vector derived by Algorithm 1. If $\xi(c^{ver}) < \eta(a^*)$, then,
(i) a^* is an optimal estimate of a^0;
(ii) the worst-case parametric error of a^* is given by

$$EP(a^*) = \overline{\sigma}(\Phi_\varsigma)\sqrt{\varepsilon^2 - \left\|\delta(a^*)\right\|_2^2} \tag{16}$$

where $\varsigma = \operatorname{supp}(a^*)$ and $\overline{\sigma}(\Phi_\varsigma)$ is the maximum singular value of Φ_ς.

Proof. See [25]. ■

Theorem 2 shows that the presented ℓ_1-relaxed-greedy algorithm is able to perform exact basis function selection, i.e. to select, within a "large" set of basis functions, the ones defining the unknown function f_0. It must be remarked that exact selection is here performed for a finite number of data. On the contrary, in standard system identification, a similar result can only be obtained when the number of data tends to infinity [22], [23], [24]. Besides exact basis function selection, the algorithm also provides an optimal parameter estimate.

Note that the optimality notion (15) is stronger than the "standard" worst-case optimality notion. Indeed, the "standard" worst-case estimation error is defined as $\overline{EP}(\widehat{a}) \doteq \sup_{a \in \overline{FPS}} \left\|a - \widehat{a}\right\|_2$, where $\overline{FPS} \doteq \{a \in \mathbb{R}^n : \left\|\widetilde{y} - \Phi a\right\|_2 \leq \varepsilon\}$, [27]. The "standard" optimal estimate is consequently defined as an estimate a^{sc} such that

$$\overline{EP}(a^{sc}) = \inf_{\widehat{a} \in \mathbb{R}^n} \sup_{a \in \overline{FPS}} \left\|a - \widehat{a}\right\|_2 . \tag{17}$$

Since $a^0 \in FPS \subseteq \overline{FPS}$, it follows that $EP(a^c) \leq \overline{EP}(a^{sc})$, showing that an optimal estimate a^c has better estimation accuracy (in a worst-case sense) with respect to a "standard" optimal estimate. Note also that the classical least squares estimate

$$a^{ls} = \arg\min_{a \in \mathbb{R}^n} \left\|\widetilde{y} - \Phi a\right\|_2 \tag{18}$$

is a "standard" optimal estimate of a^0, [27], and thus $EP(a^c) \leq \overline{EP}\left(a^{ls}\right)$.

VI. Example: Identification of vehicle lateral dynamics

In recent years, vehicle lateral dynamics control has become of great importance in the automotive field [36], [37], [38], [39], [40]. Indeed, the use of effective control systems may allow high vehicle performances in terms of safety, comfort and handling. The design of such control systems requires to have mathematical models of vehicle lateral dynamics that are accurate in describing the nonlinear vehicle behavior, and simple, in order to allow a not too difficult design.

In this example, model identification of vehicle lateral dynamics has been performed using the approach presented in Sections IV and V. This approach indeed allows the identification of models with "low" complexity, which also ensure a certain level of accuracy.

The following single-track model with 2 degrees of freedom (see e.g. [37], [39]) has been considered for the vehicle lateral dynamics:

$$\begin{aligned} \dot{\beta}(t) = & -\dot{\psi}(t) - \tfrac{c_f + c_r}{m}\tfrac{\beta(t)}{v(t)} + \tfrac{c_f}{m}\tfrac{\alpha_S(t)}{v(t)} \\ & + \tfrac{l_f c_f + l_r c_r}{m}\tfrac{\dot{\psi}(t)}{v^2(t)} + w_1(t) \\ \ddot{\psi}(t) = & -\tfrac{l_f c_f - l_r c_r}{J}\beta(t) - \tfrac{l_f^2 c_f + l_r^2 c_r}{J}\tfrac{\dot{\psi}(t)}{v(t)} + w_2(t) \end{aligned} \tag{19}$$

where $\beta(t)$ is the side-slip angle, $\dot{\psi}(t)$ is the yaw rate, $v(t)$ is the longitudinal speed, $\alpha_S(t)$ is the steering angle, $w_1(t)$ and $w_2(t)$ are noises, m is the vehicle mass, J is the momentum of inertia around the vertical axis, l_f and l_r are the front and rear distances between wheel and center of gravity, and c_f and c_r are the front and rear axle cornering stiffnesses. The following parameter values have been assumed: $m = 1798\ kg$, $J = 2900\ kgm^2$, $l_f = 0.3\ m$, $l_r = 0.5\ m$, $c_f = 76515\ Nm/rad$, and $c_r = 96540\ Nm/rad$.

A discrete-time model has been obtained by explicit Euler discretization of equations (19). This model, used in this example as the unknown "true" system to identify, is described by the following nonlinear regression equation:

$$\begin{aligned} \widetilde{y}_k = & \ 2\widetilde{y}_{k-1} - 1.087\widetilde{y}_{k-2} - 1.070\widetilde{y}_{k-1}\widetilde{p}_{k-1} \\ & -9.625\widetilde{y}_{k-1}\widetilde{p}_{k-2} + 10.69\widetilde{y}_{k-2}\widetilde{p}_{k-2} \\ & -11.52\widetilde{y}_{k-2}\widetilde{p}_{k-2}^2 + 3.715\widetilde{u}_{k-2}\widetilde{p}_{k-2} + d_k \end{aligned} \tag{20}$$

where $\widetilde{u}_k = \alpha_S(T_s k)$ and $\widetilde{p}_k = 1/v(T_s k)$ are the measured inputs, $\widetilde{y}_k = \dot{\psi}(T_s k)$ is the measured output, d_k is an unknown white noise accounting for $w_1(t)$ and $w_2(t)$, $T_s = 0.1s$, and $k = 1, 2, \ldots$

The following input signals have been considered: the steering angle $\widetilde{u}_k$ has been simulated as a white noise filtered to a maximum band of 2 rad/s. The longitudinal velocity $v(T_s k)$ has been taken as the sum of 3 sinusoids spread over the band $[0.005, 2]$ rad/s, taking values between 5 m/s and 60 m/s. d_k has been generated as a white noise. Several noise amplitudes have been used, giving noise-to-signal ratios (measured in ℓ_2 norm) ranging from 2% to 32%. For each amplitude, a set of $L = 2000$ data has been generated from the "true" system (20) and two models have

R	2%	4%	6%	8%	12%	17%	25%	32%
N_{iter}	2	3	3	3	3	4	4	4
N_e	0	0	0	0	0	0	0	0
$\overline{N}_e$	0	0	0	0	1	2	3	3
$\left\Vert a^0 - a^* \right\Vert_2$	0.017	0.043	0.025	0.086	0.090	0.356	0.101	0.354
$\left\Vert a^0 - a^{ls} \right\Vert_2$	0.057	0.131	0.143	0.236	0.594	0.653	2.109	0.898

TABLE I

IDENTIFICATION RESULTS.

been identified from these data: a sparse model obtained using the approach of Sections IV and V, and a model obtained by standard least-squares.

Identification of the two models has been performed assuming that the "true" system (20) is not known. The only information used is that this system can be described by some polynomial regression function. Note that this situation is quite realistic, since in practical applications, the exact functional structure of the system is seldom known, but some physical information on its general form is often available.

A set of $n = 22$ polynomial basis functions has been considered and the corresponding matrix $\Phi = (\Phi_1(\widetilde{x}), \ldots, \Phi_{2000}(\widetilde{x}))$ has been obtained according to

$$\begin{gathered}\Phi_k(\widetilde{x}) = [\phi_1(\widetilde{x}_k), \ldots, \phi_n(\widetilde{x}_k)] \\ = [1,\ \widetilde{y}_{k-1},\ \widetilde{y}_{k-2},\ \widetilde{u}_{k-1},\ \widetilde{u}_{k-2},\ \widetilde{p}_{k-1}, \\ \widetilde{p}_{k-2},\ \widetilde{u}_k,\ \widetilde{p}_k,\ \widetilde{y}_{k-3},\ \widetilde{u}_{k-3},\ \widetilde{p}_{k-3}, \\ \widetilde{y}_{k-1}\widetilde{p}_{k-1},\ \widetilde{y}_{k-2}\widetilde{p}_{k-1},\ \widetilde{u}_{k-2}\widetilde{p}_{k-1}, \\ \widetilde{y}_{k-1}\widetilde{p}_{k-2},\ \widetilde{y}_{k-2}\widetilde{p}_{k-2},\ \widetilde{u}_{k-2}\widetilde{p}_{k-2}, \\ \widetilde{y}_{k-2}\widetilde{p}^2_{k-2},\ \widetilde{u}_{k-2}\widetilde{p}^2_{k-2},\ \widetilde{y}_{k-3}\widetilde{p}^2_{k-2},\ \widetilde{u}_{k-3}\widetilde{p}^2_{k-2}]\end{gathered}$$

where $\widetilde{x}_k = (\widetilde{y}_{k-1},\ \widetilde{u}_{k-1},\ \widetilde{p}_{k-1},\ \ldots,\ \widetilde{y}_{k-3},\ \widetilde{u}_{k-3},\ \widetilde{p}_{k-3})$ and $k = 1, 2, \ldots, 2000$. Note that the basis function set contains the 7 functions defining the "true" system equation (20) and other 15 polynomial functions which do not appear in (20). With this choice of the basis functions, the "true" parameter vector is given by $a^0 = (0,\ 2,\ -1.087,\ 0,\ \ldots,\ 0,\ -1.070,\ 0,\ 0,\ -9.625,\ 10.69,\ 3.715,\ -11.52,\ 0,\ 0,\ 0)$.

The columns of the matrix Φ have been normalized. A sparse model with coefficient vector a^* has been identified using Algorithm 1, and its sparsity level has been evaluated by means of Theorem 1. For comparison, another model has been derived using standard least squares. The parameter vector a^{ls} of this model has been identified by means of the optimization problem (18). This problem and all those in Algorithm 1 have been solved using CVX, a package for specifying and solving convex programs [41], [42].

The identification results are shown in Table I in function of the noise-to-signal ratio R. N_{iter} is the number of iterations performed by Algorithm 1. N_e is the number of exceeding non-zero elements of a^* with respect to the "true" parameter vector a^0. $\overline{N}_e$ is the bound on N_e provided by Theorem 1. $\left\Vert a^0 - a^* \right\Vert_2$ and $\left\Vert a^0 - a^{ls} \right\Vert_2$ are the parametric errors of the sparse and least-squares models, respectively.

From these results, it can be noted that the identification Algorithm 1 is able to select exactly the "true" basis functions even in the presence of very large noise. The condition (12) ensuring that a^* is maximally sparse is satisfied even for large noise (up to a noise-to-signal ratio of 8%), indicating that this condition not only provides a theoretical motivation for using Algorithm 1, but can also be used to evaluate the sparsity of a model in practical situations. The fact that Algorithm 1 exactly selects the "true" basis functions and gives an optimal estimate according to (15) leads to a high identification accuracy. Indeed, the parametric error of a^* is significantly smaller that the parametric error of the least-square estimate a^{ls}, which satisfies the weaker optimality criterion (17).

Note also that regressors of order 3 have been included in the basis function set. The elements of a^* corresponding to these regressors have been identified as null, suggesting that Algorithm 1 could be effectively used to solve problems of model order selection.

VII. CONCLUSIONS

An algorithm for sparse identification of nonlinear functions has been proposed. A condition has been derived, ensuring that the solution identified by the algorithm is sparsest. A bound on the sparsity level of the algorithm solution with respect to the sparsest solution has been derived, useful when this condition is not satisfied. Then, a Set Membership optimality analysis has been carried out, showing that the algorithm is able to perform exact basis function selection and to provide optimal estimates.

The main advances with respect to the existing literature given in the paper are the following: 1) The condition provided for verifying that a solution is sparsest can be easily verified for any choice of the basis functions. On the contrary, the conditions available in the literature can be used in practice only for particular choices of the basis functions. 2) The exact basis function selection is performed for a finite number of data, whereas in standard system identification, a similar result can be obtained only for an infinite number of data.

REFERENCES

[1] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289 –1306, apr. 2006.

[2] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, no. 2, pp. 489 – 509, feb. 2006.

[3] E. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5406 –5425, dec. 2006.

[4] Z. Liu, S. Lin, and M. Tan, "Sparse support vector machines with l_p penalty for biomarker identification," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 7, no. 1, pp. 100 –107, jan. 2010.

[5] N. Ozay, M. Sznaier, and O. Camps, "Sequential sparsification for change detection," jun. 2008, pp. 1 –6.

[6] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *Signal Processing, IEEE Transactions on*, vol. 51, no. 1, pp. 101 – 111, jan. 2003.

[7] P. Frossard, P. Vandergheynst, R. Figueras i Ventura, and M. Kunt, "A posteriori quantization of progressive matching pursuit streams," *Signal Processing, IEEE Transactions on*, vol. 52, no. 2, pp. 525 – 535, feb. 2004.

[8] J. Tropp, J. Laska, M. Duarte, J. Romberg, and R. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *Information Theory, IEEE Transactions on*, vol. 56, no. 1, pp. 520 –544, jan. 2010.

[9] M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computations*, vol. 13, no. 4, pp. 863–882, 2001.

[10] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[11] A. Miller, *Subset selection in regression*. London, UK: Chapman and Hall, 2002.

[12] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

[13] F. Girosi., "An equivalence between sparse approximation and support vecto machines," *Neural Computation*, vol. 10, no. 6, pp. 1455–1480, 1998.

[14] N. Ozay, M. Sznaier, C. Lagoa, and O. Camps, "A sparsification approach to set membership identification of a class of affine hybrid systems," dec. 2008, pp. 123 –130.

[15] C. Feng, C. M. Lagoa, and M. Sznaier, "Hybrid system identification via sparse polynomial optimization," jun. 2010, pp. 160 –165.

[16] Y. Chen, Y. Gu, and A. Hero, "Sparse lms for system identification," apr. 2009, pp. 3125–3128.

[17] J. Tropp, "Greed is good: algorithmic results for sparse approximation," *Information Theory, IEEE Transactions on*, vol. 50, no. 10, pp. 2231 – 2242, oct. 2004.

[18] J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *Information Theory, IEEE Transactions on*, vol. 51, no. 10, pp. 3601 –3608, oct. 2005.

[19] J. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 3, pp. 1030 –1051, mar. 2006.

[20] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 6 – 18, jan. 2006.

[21] R. Gribonval, R. Figueras i Ventura, and P. Vandergheynst, "A simple test to check the optimality of a sparse signal approximation," *Signal processing*, vol. 86, no. 3, pp. 496–510, 2006.

[22] T. Söderström and P. Stoica, *System Identification*. Prentice Hall, Upper Saddle River, N.J., 1989.

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

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

[25] C. Novara, "Sparse identification of nonlinear functions and parametric set membership optimality analysis," in *Politecnico di Torino internal report*, Italy, 2011.

[26] M. Milanese and A. Vicino, "Optimal algorithms estimation theory for dynamic systems with set membership uncertainty: an overview," *Automatica*, vol. 27, pp. 997–1009, 1991.

[27] M. Milanese, "Properties of least squares in Set Memebership identification," *Automatica*, vol. 31, no. 2, pp. 327–332, 1995.

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

[29] F. Schweppe, *Uncertain dynamic systems*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[30] A. Kurzhanski and V. Veliov, *Modeling techniques for uncertain systems*. Birkhäuser, 1994.

[31] J. R. Partington, *Interpolation, Identification and Sampling*. New York: Clarendon Press - Oxford, 1997, vol. 17.

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

[33] M. Milanese and C. Novara, "Set membership identification of nonlinear systems," *Automatica*, vol. 40/6, pp. 957–975, 2004.

[34] N. Ramdani, N. Meslem, T. Ra, and Y. Candau, "Set-membership identification of continuous-time systems," in *14th IFAC Symposium on System Identification SYSID*, Newcastle, Australia, 2006.

[35] M. Sznaier, M. Wenjing, O. Camps, and L. Hwasup, "Risk adjusted set membership identification of wiener systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 5, pp. 1147–1152, 2009.

[36] A. Zanten, R. Erhardt, and G. Pfaff, "Vdc, the vehicle dynamics control system of bosch," in *SAE Technical Paper No. 950759*, 1995.

[37] J. Ackermann, "Robust control prevents car skidding," *Control Systems Magazine, IEEE*, vol. 17, no. 3, pp. 23 –31, jun. 1997.

[38] S. Suryanarayanan, M. Tomizuka, and T. Suzuki, "Design of simultaneously stabilizing controllers and its application to fault-tolerant lane-keeping controller design for automated vehicles," *Control Systems Technology, IEEE Transactions on*, vol. 12, no. 3, pp. 329 – 339, may. 2004.

[39] R. Rajamani, *Vehicle Dynamics and Control*. Springer Verlag, 2006.

[40] V. Cerone, M. Milanese, and D. Regruto, "Yaw stability control design through a mixed-sensitivity approach," *Control Systems Technology, IEEE Transactions on*, vol. 17, no. 5, pp. 1096 –1104, sep. 2009.

[41] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, Aug. 2010.

[42] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/ boyd/graph_dcp.html.