Available online at www.sciencedirect.com

SCIENCE DIRECT®

Automatica 40 (2004) 957–975

automatica

www.elsevier.com/locate/automatica

Set Membership identification of nonlinear systems☆

Mario Milanese*, Carlo Novara

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Received 27 November 2002; received in revised form 21 January 2004; accepted 5 February 2004

Abstract

In the paper the problem of identifying nonlinear dynamic systems, described in nonlinear regression form, is considered, using finite and noise-corrupted measurements. Most methods in the literature are based on the estimation of a model within a finitely parametrized model class describing the functional form of involved nonlinearities. A key problem in these methods is the proper choice of the model class, typically realized by a search, from the simplest to more complex ones (linear, bilinear, polynomial, neural networks, etc.). In this paper an alternative approach, based on a Set Membership framework is presented, not requiring assumptions on the functional form of the regression function describing the relations between measured input and output, but assuming only some information on its regularity, given by bounds on its gradient. In this way, the problem of considering approximate functional forms is circumvented. Moreover, noise is assumed to be bounded, in contrast with statistical methods, which rely on assumptions such as stationarity, ergodicity, uncorrelation, type of distribution, etc., whose validity may be difficult to test reliably and is lost in presence of approximate modeling. In this paper, necessary and sufficient conditions are given for the validation of the considered assumptions. An optimal interval estimate of the regression function is obtained, providing its uncertainty range for any assigned regressor values. The set estimate allows to derive an optimal identification algorithm, giving estimates with minimal guaranteed L_p error on the assigned domain of the regressors. The properties of the optimal estimate are investigated and its worst-case L_p identification error is evaluated. The presented approach is tested and compared with other nonlinear methods on the identification of a water heater, a mechanical system with input saturation and a vehicle with controlled suspensions.

Keywords: System identification; Nonlinear systems; Identification algorithms; Uncertain dynamic systems; Modeling errors; Model approximation; Set estimate; Semi-active suspensions

1. Introduction

Consider a nonlinear discrete time dynamic system, described in the regression form

$$y^{t+1} = f_o(w^t), \tag{1}$$

where $w^t = [y^t \dots y^{t-n_y+1} u_1^t \dots u_1^{t-n_1+1} \dots u_m^t \dots u_m^{t-n_m+1}]$ and $y^t, u_1^t, \dots, u_m^t \in \mathbb{R}$, $f_o : \mathbb{R}^n \to \mathbb{R}$, $n = n_y + \sum_{i=1}^m n_i$.

Suppose that the function f_o is not known, but a set of noise corrupted measurements $\tilde{y}^t$ and $\tilde{w}^t$ of y^t and w^t, $t = 1, 2, \dots, T$ is available, and it is of interest to make an inference on the system (e.g. identification, prediction, smoothing, filtering, control design, decision making, fault detection, etc.). In this paper the focus is on the case that the desired inference is identification of f_o. The case that desired inference is prediction, has been considered in Novara and Milanese (2001) and Milanese and Novara (2002).

In the identification problem investigated here, the aim is to find an estimate $\hat{f}$ of f_o giving small, possibly minimal, identification error $f_o - \hat{f}$. However, this error is not known and, since data are finite and noise corrupted, no reliable estimate on the identification error can be derived if no information is available on f_o and on noise. The information on f_o is typically given by assuming that it belongs to some subset $\mathscr{F}$ of functions. In some cases, the knowledge of the laws governing the system (mechanical, economical, biological,

☆ This research was supported in part by Ministero dell'Universitá e della Ricerca Scientifica e Tecnologica under the Project "Robustness techniques for control of uncertain systems" and by GM-FIAT World Wide Purchasing Italia under the Project "Modeling of vehicles vertical accelerations". This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Brett Ninness under the Director of Editor Torsten Söderström.

* Corresponding author. Tel.: +39-011-5647020; fax: +39-011-5647099.

E-mail addresses: mario.milanese@polito.it (M. Milanese), carlo.novara@polito.it (C. Novara).

0005-1098/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2004.02.002

etc.) generating the data, may allow to have reliable information on its structure. In many other situations, due to the fact that the laws are too complex or not sufficiently known, this is not possible or not convenient and the usual approach is to consider that f_o belongs to a finitely parametrized set of functions $\mathscr{F}(\theta) \doteq \{f(w,\theta)=\sum_{i=1}^{r} \alpha_i\sigma_i(w,\beta_i), \beta_i \in \mathbb{R}^q\}$, where $\theta=[\alpha,\beta]$ and the σ_i's are given functions. Then, measured data are used to derive an estimate $\hat{\theta}$ of θ and $f(w,\hat{\theta})$ is used as estimate of f_o. Basic to this approach is the proper choice of the parametric family of functions $f(w,\theta)$, typically realized by some search on different functional forms of the σ_i's, e.g. linear, polynomial, sigmoidal, wavelet, etc. and on the number r, (Sjöberg et al., 1995). This search may be quite time consuming, and in any case leads to approximate model structures only. The evaluation of the effects of such approximation on identification errors appears at present to be a largely open problem. Another critical point is related to the fact that the estimate $\hat{p}$ of p are usually obtained by a prediction error method, which requires the minimization of the error function

$$V(\theta,\Phi_T) = \frac{1}{T}\sum_{t=0}^{T-1} |\tilde{y}^{\,t+1} - f(\varphi^t,\theta)|^2, \tag{2}$$

where φ^t is a regression vector and $\Phi_T = [\varphi^0, \varphi^1, \ldots, \varphi^{\mathrm{T}}]$. Several choices can be taken for the regressor φ^t. Widely used are the following ones:

$$\varphi^t = \tilde{w}^t = [\tilde{y}^{\,t} \ldots \tilde{y}^{\,t-n_y+1}$$
$$\tilde{u}_1^t \ldots \tilde{u}_1^{t-n_1+1} \ldots \tilde{u}_m^t \ldots \tilde{u}_m^{t-n_m+1}]$$
$$\varphi^t = \hat{w}^t = [f(\hat{w}^{t-1},p) \ldots f(\hat{w}^{t-n_y},p)$$
$$\tilde{u}_1^t \ldots \tilde{u}_1^{t-n_1+1} \ldots \tilde{u}_m^t \ldots \tilde{u}_m^{t-n_m+1}]$$

leading to NARX and NOE models, respectively, (see e.g. Sjöberg et al., 1995). Such an approach is often indicated as prediction error (PE) method, since $V(\theta,\Phi_T)$ is an estimate of the prediction error for the given regressor choice. The functional $V(\theta,\Phi_T)$ is convex w.r.t. θ only if the basis functions σ_i are not dependent on the tunable parameters, i.e. $\theta=[\alpha]$, and ARX regression structure is chosen, as it happens e.g. for NARX polynomial models. However, it is well known that fixed basis functions suffers of the "curse of dimensionality", i.e. the number r of terms required for obtaining a given approximation increases exponentially with the dimension n of the regressor space, while basis functions σ_i dependent on the tunable parameters, such as wavelets or neural networks, have much powerful approximation properties, requiring only polynomial growth (Barron, 1993; Hornik, Stinchcombe, White, & Auer, 1994). Unfortunately, with such basis functions, $V(\theta,\Phi_T)$ is no more convex w.r.t. θ, even for an ARX regression structure, giving rise to possible deteriorations in approximation, due to trapping in local minima during its minimizations. Other problems arise in giving a measure of identification error $f_o(w) - f(w,\hat{\theta})$. Under the standard assumption that noise affecting measurements is a stochastic process, the quality of identification is usually measured by the variance of this error. However, no reliable finite sample results on the estimate of this variance are available. Moreover, in case of approximate model class, where $f_o(w) \notin \mathscr{F}(\theta)$, a bias term is present, whose reliable evaluation is also difficult.

In order to circumvent such problems, in this paper an alternative approach is taken, formulating the problem in a Set Membership (SM) framework, used in linear systems identification to deal with approximate model structures and finite sample accuracy evaluation, see e.g. Milanese and Tempo (1985); Milanese and Vicino (1991); Milanese, Norton, Piet Lahanier, and Walter (1996); Partington (1997) and Chen and Gu (2000). No assumptions on the functional form of f_o is required, and an assumption on its regularity is used instead, given by bounds on the gradient of f_o. An optimal estimate of f_o, having minimal guaranteed L_p identification error is derived, not requiring iterative minimization and thus avoiding trapping in local minima. The optimal estimate is derived evaluating tight bounds on f_o. These bounds give a measure of achieved accuracy in evaluating f_o, which can be useful for successive robust analysis or design using the identified model, e.g. for guaranteed stability analysis of errors in simulation for future inputs (Sontag, 1992; Milanese & Novara, 2003) or for robust control design (Freeman & Kokotovic, 1996; Qu, 1998).

It can be noted that the proposed approach has strong connections with method used for approximation, interpolation or optimization of multivariable functions with bounded derivatives, from the knowledge of a finite number of their values (see e.g. Traub, Wasilkowski, & Woźniakowski, 1988; Novak, 1988; Wasilkowski & Woźniakowski, 2001 and the references therein). In this literature, noise free measurements are typically assumed, and weaker optimality concepts are considered than the one of the present paper (see the remark at the end of next section for a more specific discussion).

The paper is organized as follows. In Section 2 the identification problem is formulated in a SM framework, defining the type of assumptions considered, the guaranteed identification error and optimality concept. In Section 3, necessary and sufficient conditions are given for assumptions validation (intended as consistency of assumptions with measured data) and it is shown how they can be used for assessing the constants appearing in the assumptions. Also, tight lower and upper bounds $\underline{f}(w)$ and $\overline{f}(w)$ of $f_o(w)$ are derived. In Section 4, Hyperbolic Voronoi Diagrams are introduced and used to investigate the properties of the bounds $\underline{f}(w)$ and $\overline{f}(w)$. In Section 5, an optimal point estimate of f_o, having minimal guaranteed L_p identification error, is obtained and its properties are investigated. In Section 6, two variations of the method are proposed, which may give significant improvement, allowing adaption to properties of data, such as variable gradient bounds and quite different magnitude of gradient components. In Section 7, the overall identification procedure is summarized, indicating step-by-step the operations to be performed. In Section 8, the presented method is tested and compared with other nonlinear methods on the identification of a water heater, a mechanical system with input saturation and of a vehicle with controlled suspension.

2. The nonlinear SM approach

Consider that sets of noise corrupted data $\tilde{Y}^{\mathrm{T}} = [\tilde{y}^{\,2}, \tilde{y}^{\,3}, \ldots, \tilde{y}^{\,\mathrm{T}+1}]$ and $\tilde{W}^{\mathrm{T}} = [\tilde{w}^1, \tilde{w}^2, \ldots, \tilde{w}^{\mathrm{T}}]$ generated by (1) are available. Then

$$\tilde{y}^{\,t+1} = f_o(\tilde{w}^t) + d^t, t = 1, 2, \ldots, T, \tag{3}$$

where the term d^t accounts for the fact y^{t+1} and w^t are not exactly known, a setting often indicated in the literature as error-in-variables.

The aim is to derive an estimate $\hat{f}$ of f_o from available measurements $(\tilde{Y}^{\mathrm{T}}, \tilde{W}^{\mathrm{T}})$, i.e. $\hat{f} = \phi(\tilde{Y}^{\mathrm{T}}, \tilde{W}^{\mathrm{T}})$. The operator ϕ, called identification algorithm, should be chosen to give small (possibly minimal) $L_p(W)$ error $e(\hat{f}) = \|f_o - \hat{f}\|_p$, where W is a given bounded set in $\mathbb{R}^n$ and $\|f\|_p \doteq \left[\int_W |f(w)|^p \,\mathrm{d}w\right]^{1/p}$, $p \in [1, \infty)$ and $\|f\|_\infty \doteq \text{ess-sup}_{w \in W} |f(w)|$. This error is not known, since from available data it is only known that $f_o \in \mathscr{F}(\tilde{Y}^{\mathrm{T}}, \tilde{W}^{\mathrm{T}})$, the set of all f that could have generated the data. If no assumptions is made on f_o, this set, even in case of exact measurements, is unbounded, since the mapping generating data from f is not injective, i.e. infinitely many f measured at $\tilde{w}^t$, $t = 1, \ldots, T$ give the same values $\tilde{y}^{\,t+1}$, $t = 1, \ldots, T$. Then, whatever algorithm ϕ is chosen, no information on the identification error can be derived, unless some assumptions are made on the function f_o and the noise d. The typical approach in the literature is to assume a finitely parametrized functional form for f_o (linear, bilinear, neural network, etc.) and statistical models on the noise (Sjöberg et al., 1995). In the present SM approach, different and somewhat weaker assumptions are taken, not requiring the selection of a functional form for f_o, but related to its rate of variation. Moreover, the noise sequence $D^{\mathrm{T}} = [d^1, d^2, \ldots, d^{\mathrm{T}}]$ is only supposed to be bounded.

Prior assumptions on f_o:

$$f_o \in \mathscr{F}(\gamma) \doteq \{f \in C^1(W) : \|f'(w)\| \leqslant \gamma, \forall w \in W\}.$$

Prior assumptions on noise:

$$D^{\mathrm{T}} \in \mathscr{D} \doteq \{[d^1, d^2, \ldots, d^{\mathrm{T}}] : |d^t| \leqslant \varepsilon^t, t = 1, 2, \ldots, T\}.$$

Here, $f'(w)$ denotes the gradient of $f(w)$ and $\|x\| \doteq \sqrt{\sum_{i=1}^n x_i^2}$ is the Euclidean norm.

A key role in this Set Membership framework is played by the Feasible Systems Set, often called "unfalsified systems set" , i.e. the set of all systems consistent with prior information and measured data.

Definition 1 (*Feasible Systems Set*). The Feasible Systems Set FSS^{T} is

$$FSS^{\mathrm{T}} \doteq \{f \in \mathscr{F}(\gamma) : |\tilde{y}^{\,t+1} - f(\tilde{w}^t)| \leqslant \varepsilon^t, \quad t = 1, 2, \ldots, T\}. \tag{4}$$

The feasible systems set FSS^{T} summarizes all the information on the mechanism generating the data that is available up to time T. If prior assumptions are "true" , then $f_o \in FSS^{\mathrm{T}}$, which is an important property for evaluating the accuracy of inferences that can be done on the system. In particular, it follows that $f_o(w)$ is bounded as

$$\underline{f}(w) \leqslant f_o(w) \leqslant \bar{f}(w), \forall w \in W, \tag{5}$$

where

$$\bar{f}(w) = \sup_{f \in FSS^{\mathrm{T}}} f(w),$$

$$\underline{f}(w) = \inf_{f \in FSS^{\mathrm{T}}} f(w). \tag{6}$$

Provided that the prior assumptions hold, $\bar{f}(w)$ and $\underline{f}(w)$ are the tightest upper and lower bounds of $f_o(w)$ and are called *optimal bounds*.

As typical in any identification theory, the problem of checking the validity of prior assumptions arises. The only thing that can be actually done is to check if prior assumptions are invalidated by the data, evaluating if no system exists consistent with data and assumptions, i.e. if FSS^{T} is empty. However, it is usual to introduce the concept of prior assumption validation as consistency with the measured data, i.e. FSS^{T} not empty (Milanese et al., 1996; Chen & Gu, 2000).

Definition 2 (*Validation of prior assumptions*). Prior assumptions are considered validated if $FSS^{\mathrm{T}} \neq \emptyset$.

Note that the fact that the prior assumptions are validated, i.e. are consistent with the present data, does not exclude that they may be invalidated by future data (Popper, 1969). In the following, the FSS^{T} is assumed to be non-empty. If empty, the prior assumptions on the system and the noise are invalidated by data and have to be suitably modified to give a non-empty FSS^{T} as discussed in Section 3.

An identification algorithm ϕ is an operator mapping all available information about function f_o, noise d, data $(\tilde{Y}^{\mathrm{T}}, \tilde{W}^{\mathrm{T}})$ until time T, summarized by FSS^{T}, into an estimate $\hat{f}$ of f_o:

$$\phi(FSS^{\mathrm{T}}) = \hat{f} \simeq f_o.$$

The related L_p error is:

$$e(\hat{f}) = e(\phi(FSS^{\mathrm{T}})) = \|f_o - \hat{f}\|_p.$$

This error cannot be exactly computed, since it is only known that $f_o \in FSS^{\mathrm{T}}$, but its tightest bound is given by $e(\hat{f}) \leqslant \sup_{f \in FSS^{\mathrm{T}}} \|f - \hat{f}\|_p$. This motivates the following definition of the identification error, often indicated as local worst-case or guaranteed error.

Definition 3 ((*Local*) *identification error*). The (local) identification error of $\hat{f} = \phi(FSS^{\mathrm{T}})$ is

$$E[\phi(FSS^{\mathrm{T}})] = E(\hat{f}) \doteq \sup_{f \in FSS^{\mathrm{T}}} \|f - \hat{f}\|_p.$$

Looking for algorithms that minimize this identification error, leads to the following optimality concepts.

Definition 4 ((*Locally*) *optimal algorithm*). An algorithm ϕ^* is called (locally) optimal if

$$\begin{aligned} E[\phi^*(FSS^{\mathrm{T}})] &= \inf_{\phi} E[\phi(FSS^{\mathrm{T}})] \\ &= \inf_{\hat{f}} \sup_{f \in FSS^{\mathrm{T}}} \|f - \hat{f}\|_p = r_I. \end{aligned}$$

Note that an optimal algorithm, if it exists, may give an optimal estimate $f^* = \phi^*(FSS^{\mathrm{T}})$ not necessarily in FSS^{T}. The quantity r_I, called (local) radius of information, gives the minimal identification error that can be guaranteed by any estimate based on the available information up to time T.

Remark. The (local) identification error actually depends on f_o and D^{T}, i.e. $E(\hat{f}) = E(\hat{f}, f_o, D^{\mathrm{T}})$. A global identification error of given algorithms ϕ is often considered, defined as:

$$E^{\mathrm{g}}(\phi) \doteq \sup_{\substack{f_o \in \mathscr{F}(\gamma) \\ D^{\mathrm{T}} \in \mathscr{D}}} E(\phi(FSS^{\mathrm{T}}), f_o, D^{\mathrm{T}}).$$

An algorithm ϕ^{g} is called globally optimal if $E^{\mathrm{g}}(\phi^{\mathrm{g}}) = \inf_{\phi} E^{\mathrm{g}}(\phi)$. This is the optimality concept usually investigated in the approximation theory literature (Traub et al., 1988; Novak, 1988; Wasilkowski & Woźniakowski, 2001). Note that a (locally) optimal algorithm ϕ^* is globally optimal, but ϕ^{g} is not in general locally optimal. Thus, the (local) optimality concept investigated in this paper is stronger than the global optimality concept investigated in the above cited literature. As just noted before, for given algorithm, the tightest bound on $\|f_o - \phi(FSS^{\mathrm{T}})\|$ is given by $E(\phi(FSS^{\mathrm{T}}), f_o, D^{\mathrm{T}})$. The algorithm ϕ^* minimizes $E(\phi(FSS^{\mathrm{T}}), f_o, D^{\mathrm{T}})$ for any f_o and D^{T}, while ϕ^{g} minimizes it only for worst case $f \in \mathscr{F}(\gamma)$ and noise sequence in D. Then the ratio:

$$\frac{E(\phi^{\mathrm{g}}(FSS^{\mathrm{T}}), f_o, D^{\mathrm{T}})}{E(\phi^*(FSS^{\mathrm{T}}), f_o, D^{\mathrm{T}})} \doteq \alpha(\phi^{\mathrm{g}}(FSS^{\mathrm{T}}), f_o, D^{\mathrm{T}})$$

give a measure of the degradation in the guaranteed identification error of using ϕ^{g} instead of ϕ^*, for the given f_o and noise realization D^{T}. Clearly, $\alpha \geqslant 1$, indicating that ϕ^{g} cannot be better than ϕ^*. Indeed, relatively simple globally optimal algorithms ϕ_L^{g} exist (e.g. linear in the measured values) for which $\alpha(\phi_L^{\mathrm{g}}, f_o, D^{\mathrm{T}})$ may be arbitrarily large (see Traub et al., 1988). Thus, globally optimal algorithms may lead to estimates with large degradation with respect locally optimal estimates. This motivates the interest for investigating (locally) optimal algorithms ϕ^*. In the rest of the paper the local optimality concept will be considered and the term (local) will be omitted.

3. Assumptions validation and optimal bounds evaluation

Necessary and sufficient conditions for checking the assumptions validity are now given. Let us define the functions:

$$\begin{aligned} f_u(w) &\doteq \min_{t=1,\ldots,T} (\bar{h}^t + \gamma\|w - \tilde{w}^t\|), \\ f_l(w) &\doteq \max_{t=1,\ldots,T} (\underline{h}^t - \gamma\|w - \tilde{w}^t\|), \end{aligned} \tag{7}$$

where $\bar{h}^t \doteq \tilde{y}^{\,t+1} + \varepsilon^t$ and $\underline{h}^t \doteq \tilde{y}^{\,t+1} - \varepsilon^t$.

Theorem 1.

(i) *A necessary condition for prior assumptions to be validated is*: $f_u(\tilde{w}^t) \geqslant \underline{h}^t, t = 1, 2, \ldots, T$.

(ii) *A sufficient condition for prior assumptions to be validated is*: $f_u(\tilde{w}^t) > \underline{h}^t, t = 1, 2, \ldots, T$.

Proof. We have to prove that if prior assumptions are validated, i.e. $FSS^{\mathrm{T}} \neq \emptyset$, then $f_u(\tilde{w}^t) \geqslant \underline{h}^t$, $t = 1, 2, \ldots, T$.

Let $f \in C^1(W)$. From mean value theorem it follows that for every $w \in W$, and for each $t = 1, 2, \ldots, T$, a $\hat{w}^t \in W$ exists such that $f(w) = f(\tilde{w}^t) + f'(\hat{w}^t) \cdot (w - \tilde{w}^t)$. Now let $f \in FSS^{\mathrm{T}}$, then $f(w) \leqslant \tilde{y}^{\,t+1} + \varepsilon^t + \|f'(\hat{w}^t)\| \, \|w - \tilde{w}^t\| \leqslant \tilde{y}^{\,t+1} + \varepsilon^t + \gamma\|w - \tilde{w}^t\|$. This holds for $\forall w \in W$ and $t = 1, 2, \ldots, T$ then, from (7) we have

$$f_u(w) \geqslant f(w), \forall w \in W. \tag{8}$$

Similarly it can be proven that:

$$f_l(w) \leqslant f(w), \forall w \in W. \tag{9}$$

From (8), (9) it follows that $f_u(\tilde{w}^t) \geqslant f_l(\tilde{w}^t)$, $t = 1, 2, \ldots, T$ and from (7) it follows that $f_l(\tilde{w}^t) \geqslant \underline{h}^t$, $t = 1, 2, \ldots, T$, then $f_u(\tilde{w}^t) \geqslant \underline{h}^t$, $t = 1, 2, \ldots, T$.

(ii) Suppose that $f_u(\tilde{w}^t) > \underline{h}^t$, $t = 1, 2, \ldots, T$. We have to prove that $FSS^{\mathrm{T}} \neq \emptyset$, i.e. that a function $f \in \mathscr{F}(\gamma)$ can be found such that $|\tilde{y}^{\,t+1} - f(\tilde{w}^t)| \leqslant \varepsilon^t$, $t = 1, 2, \ldots, T$. For given $w \in W$, let $\bar{t}$ and $\underline{t}$ be such that $\bar{t} = \arg\min_t (\bar{h}^t + \gamma\|w - \tilde{w}^t\|)$ and $\underline{t} = \arg\max_t (\underline{h}^t - \gamma\|w - \tilde{w}^t\|)$. Thus, the following inequalities hold: $f_u(w) - f_l(w) = \bar{h}^{\bar{t}} - \underline{h}^{\underline{t}} + \gamma(\|w - \tilde{w}^{\bar{t}}\| + \|w - \tilde{w}^{\underline{t}}\|) \geqslant \bar{h}^{\bar{t}} - \underline{h}^{\underline{t}} + \gamma\|\tilde{w}^{\bar{t}} - \tilde{w}^{\underline{t}}\| \geqslant f_u(\tilde{w}^{\underline{t}}) - \underline{h}^{\underline{t}} > 0$. Since w is an arbitrary point of W, then

$$f_l(w) < f_u(w), \forall w \in W. \tag{10}$$

By defining $f_c(w) = \frac{1}{2}[f_l(w) + f_u(w)]$, this inequality implies that

$$f_l(w) < f_c(w) < f_u(w), \forall w \in W. \tag{11}$$

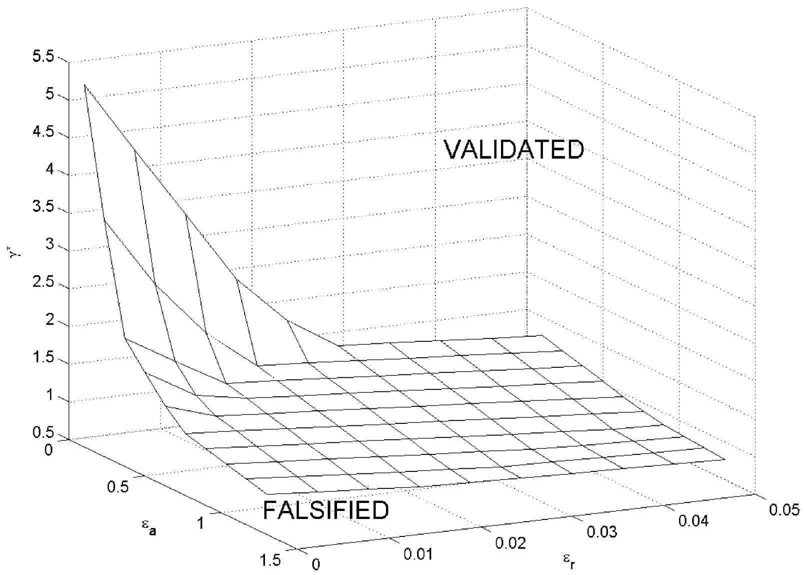

Fig. 1. Example of validation surface.

On the other hand, from (7) it follows $f_u(\tilde{w}^t) \leqslant \tilde{y}^{\,t+1} + \varepsilon^t$, $f_l(\tilde{w}^t) \geqslant \tilde{y}^{\,t+1} - \varepsilon^t$, $\forall t$, which, together with (11), implies $\tilde{y}^{\,t+1} - \varepsilon^t \leqslant f_l(\tilde{w}^t) < f_c(\tilde{w}^t) < f_u(\tilde{w}^t) \leqslant \tilde{y}^{\,t+1} + \varepsilon^t$, $\forall t$, and then $|\tilde{y}^{\,t+1} - f_c(\tilde{w}^t)| < \varepsilon^t$, $t = 1, 2, \ldots, T$, which can be expressed as

$$|\tilde{y}^{\,t+1} - f_c(\tilde{w}^t)| \leqslant \varepsilon^t - \rho, \rho > 0, \; t = 1, 2, \ldots, T. \tag{12}$$

Note that $f_c \notin FSS^{\mathrm{T}}$, since $|\tilde{y}^{\,t+1} - f_c(\tilde{w}^t)| \leqslant \varepsilon^t$, $t = 1, 2, \ldots, T$, but $f_c \notin \mathscr{F}(\gamma)$, being not everywhere differentiable, see Theorem 6. However, a function belonging to FSS^{T} can be obtained by suitably approximating f_c.

Let $\overline{\mathscr{F}(\gamma)}$ be the closure of $\mathscr{F}(\gamma)$ with respect to (wrt) the norm $\|f\|_S = \|f\|_\infty + \|f'\|_2$. The linear space $C^1(W)$ of continuously differentiable functions over W, embedded with the norm $\|\cdot\|_S$ is not a Banach space, since it is not complete. Let $F(W)$ be the Banach space obtained by completing $C^1(W)$ wrt $\|\cdot\|_S$. Given a Cauchy sequence $\{f_m\}$, $\lim_{m,l\to\infty} \|f_m - f_l\|_S = 0$ implies $\lim_{m,l\to\infty} \|f_m - f_l\|_\infty = 0$ and $\lim_{m,l\to\infty} \|f'_m - f'_l\|_2 = 0$. Since $f_m \in C^1(W)$, $\forall m$, $\|f_m\|_\infty = \sup_{w\in W} |f_m(w)|$ and then convergence of $\|f_m - f_l\|_\infty$ is uniform. On the contrary, convergence of $\|f'_m - f'_l\|_2$ is almost everywhere (a.e.) only. Then $F(W)$ is the space of continuous functions, differentiable a.e. in W and: $\overline{\mathscr{F}(\gamma)} = \{f \in F(W), \|f'(w)\| \leqslant \gamma, \text{ a.e. in } W\}$. From Theorem 6 in Section 5 it follows that $f_c \in \overline{\mathscr{F}(\gamma)}$. Since $\overline{\mathscr{F}(\gamma)}$ is the closure of $\mathscr{F}(\gamma)$ wrt the norm $\|f\|_S$, $\mathscr{F}(\gamma)$ is dense in $\overline{\mathscr{F}(\gamma)}$, i.e. $\forall \delta > 0$ a function $f_\delta \in \mathscr{F}(\gamma)$ can be found such that $\|f_c - f_\delta\|_S < \delta$. Since $\|f\|_S \geqslant \|f\|_\infty = \sup_{w\in W} |f(w)|$, $\forall f$, it follows that

$$|f_c(w) - f_\delta(w)| < \delta, \; \forall w. \tag{13}$$

Then, from (12) and (13) it results: $|\tilde{y}^{\,t+1} - f_\delta(\tilde{w}^t)| \leqslant |\tilde{y}^{\,t+1} - f_c(\tilde{w}^t)| + |f_c(\tilde{w}^t) - f_\delta(\tilde{w}^t)| \leqslant \varepsilon^t - \rho + \delta, t = 1, 2, \ldots, T$. Taking $\delta < \rho$, we have that $f_\delta \in \mathscr{F}(\gamma)$ and $|\tilde{y}^{\,t+1} - f_\delta(\tilde{w}^t)| \leqslant \varepsilon^t$, $t = 1, 2, \ldots, T$, thus proving that $FSS^{\mathrm{T}} \neq \emptyset$. □

Note that there is essentially no "gap" between the necessary and sufficient conditions, since condition $\bar{f}(\tilde{w}^t) \geqslant \underline{h}^t + \delta$, $t = 1, 2, \ldots, T$ is sufficient for any $\delta > 0$ arbitrarily small and necessary for $\delta = 0$. In the rest of paper it is assumed that the sufficient condition holds. If not, values of the constants appearing in the assumptions on function f_o and on noise d^t have to be suitably modified. The validation Theorem 1 can be used for assessing the values of such constants so that sufficient condition holds.

Let us consider a relative plus absolute model for the error bound given as

$$|d^t| \leqslant \varepsilon^t = \varepsilon_r |y^{t+1}| + \varepsilon_a, \varepsilon_r, \varepsilon_a \geqslant 0$$

In the space $(\varepsilon_r, \varepsilon_a, \gamma)$, the function

$$\gamma^*(\varepsilon_r, \varepsilon_a) \doteq \inf_{\varepsilon_r, \varepsilon_a, \gamma : FSS^{\mathrm{T}} \neq \emptyset} \gamma \tag{14}$$

individuates a surface that separate falsified values of $\varepsilon_r, \varepsilon_a$ and γ from validated ones, see Fig. 1, where the surface related to the Example 1 presented in Section 8 is shown.

Clearly, $\varepsilon_r, \varepsilon_a$ and γ must be chosen in the validated parameters region, possibly using information from the experimental setting for assessing e.g. the respective relevance of relative and absolute components of the noise. On the other hand, useful information on γ values can be obtained by deriving (e.g. from a neural networks approximation or directly from data) some estimates of $f'_o(w)$. See Section 8 and the examples for more detailed discussion on the use of this procedure for the selection of $(\varepsilon_r, \varepsilon_a, \gamma)$ values.

The functions f_u and f_l allow also to solve the problem of finding the optimal interval estimate of $f_o(w)$ for given

$w \in W$. In fact, from (5) and (6) it follows that the smallest interval guaranteed to include $f_o(w)$, is given by the interval $[\underline{f}(w), \bar{f}(w)]$, where $\underline{f}(w)$ and $\bar{f}(w)$ are the optimal bounds. The following theorem shows that the optimal bounds are actually given by f_u and f_l.

Theorem 2. *The functions* f_u and f_l *given in* (7) *are optimal bounds, i.e.*

$$\bar{f}(w) = \min_{t=1,\dots,T} (\bar{h}^t + \gamma\|w - \tilde{w}^t\|) \doteq f_u(w),$$

$$\underline{f}(w) = \max_{t=1,\dots,T} (\underline{h}^t - \gamma\|w - \tilde{w}^t\|) \doteq f_l(w). \tag{15}$$

Proof. Inequalities (8) and (9) hold for every $f \in FSS^{\mathrm{T}}$, then $f_l(w) \leqslant f_o(w) \leqslant f_u(w)$, $\forall w \in W$. These bounds on f_o imply the following bounds on noise d : $f_l(\tilde{w}^t) - \tilde{y}^{\,t+1} \leqslant \tilde{y}^{\,t+1} - d^t = f_o(\tilde{w}^t) \leqslant f_u(\tilde{w}^t) - \tilde{y}^{\,t+1}$, $t = 1, 2, \dots, T-1$, then

$$d^t \in B_d^t = \{\hat{d} : f_l(\tilde{w}^t) \leqslant \tilde{y}^{\,t+1} - \hat{d} \leqslant f_u(\tilde{w}^t)\}.$$

For given $w \in W$, let $\bar{t} = \arg\min_{t=1,\dots,T-1} (\tilde{y}^{\,t+1} + \varepsilon^t + \gamma\delta^t + \gamma\|w - \tilde{w}^t\|)$. Also, let $f \in FSS^{\mathrm{T}}$. Therefore, a $\hat{w}^{\bar{t}} \in W$ and a $d^t \in B_d^t$ exist such that $f(w) = \tilde{y}^{\,\bar{t}+1} - d^{\bar{t}} + f'(\hat{w}^{\bar{t}}) \cdot (w - \tilde{w}^{\bar{t}})$, so that

$$\begin{aligned}
&\sup_{\substack{d^t \in B_d^t, t=1,\dots,T-1 \\ \|f'(\hat{w}^t)\| \leqslant \gamma, t=1,\dots,T-1}} f(w) \\
&\quad = \sup_{\substack{d^{\bar{t}} \in B_{\bar{t}}^d \\ \|f'(\hat{w}^{\bar{t}})\| \leqslant \gamma}} [\tilde{y}^{\,\bar{t}+1} - d^{\bar{t}} + f'(\hat{w}^{\bar{t}}) \cdot (w - \tilde{w}^{\bar{t}})] \\
&\quad = \sup_{d^{\bar{t}} \in B_{\bar{t}}^d} (\tilde{y}^{\,\bar{t}+1} - d^{\bar{t}}) + \sup_{\|f'(\hat{w}^{\bar{t}})\| \leqslant \gamma} [f'(\hat{w}^{\bar{t}}) \cdot (w - \tilde{w}^{\bar{t}})] \\
&\quad = f_u(\tilde{w}^{\bar{t}}) + \gamma\|w - \tilde{w}^{\bar{t}}\|.
\end{aligned}$$

Since $f_u(\tilde{w}^{\bar{t}}) = \tilde{y}^{\,\bar{t}+1} + \varepsilon^{\bar{t}}$ and $f_u(w) = \tilde{y}^{\,\bar{t}+1} + \varepsilon^{\bar{t}} + \gamma\|w - \tilde{w}^{\bar{t}}\|$, we have that

$$\sup_{\substack{d^t \in B_d^t, t=1,\dots,N \\ \|f'(\hat{w}^t)\| \leqslant \gamma, t=1,\dots,N}} f(w) = \tilde{y}^{\,\bar{t}+1} + \varepsilon^{\bar{t}} + \gamma\|w - \tilde{w}^{\bar{t}}\| = f_u(w).$$

This holds for all $w \in W$, then $f_u(w) = \sup_{f \in FSS^{\mathrm{T}}} f(w) = \bar{f}(w)$. The proof that $f_l(w) = \inf_{f \in FSS^{\mathrm{T}}} f(w) = \underline{f}(w)$ is similar. □

4. Hyperbolic Voronoi Diagrams (HVD)

In this section the concept of hyperbolic Voronoi diagram (HVD) is introduced. The HVD are a generalization of standard Voronoi diagrams (see e.g. Edelsbrunner, 1987) and are used to investigate the properties of the optimal bounds $\underline{f}$ and $\bar{f}$ and of the optimal identification algorithm derived in the next section. The HVD are defined as follows.

Consider the set of points: $\tilde{W}^{\mathrm{T}} \doteq [\tilde{w}^1, \tilde{w}^2, \dots, \tilde{w}^{\mathrm{T}}]$ and a $T \times T$ antisymmetric matrix η. Then define:

- The $(n-1)$-dimensional hyperbola $H^{t\tau}$:

  $H^{t\tau} \doteq \{w \in \mathbb{R}^n : \|w - \tilde{w}^t\| - \|w - \tilde{w}^\tau\| = \eta^{t\tau}, t \neq \tau\}$
- The n-dimensional regions $S^{t\tau}$ containing $\tilde{w}^t$:

  $S^{t\tau} \doteq \{w \in \mathbb{R}^n : \|w - \tilde{w}^t\| - \|w - \tilde{w}^\tau\| < \eta^{\tau t}, t \neq \tau\}$
- The hyperbolic cell C^t : $C^t \doteq \bigcap_{\tau \neq t} S^{t\tau}$.

Note that some cell C^t may be empty (see Theorem 3 below). The intersections between the surfaces $H^{t\tau}$ generate other cells of dimension d, with $0 \leqslant d \leqslant n-1$ called d-faces. The cells C^t are called n-faces while the 0-faces are also called vertices.

Definition 5 (*Hyperbolic Voronoi Diagram*). The Hyperbolic Voronoi Diagram $V(\tilde{W}^{\mathrm{T}}, \eta)$ is defined as the set of all d-faces, $0 \leqslant d \leqslant n$.

If $\eta^{t\tau} = 0, \forall t, \tau$, all hyperbola $H^{t\tau}$ degenerate into hyperplanes and the definitions become the ones of standard Voronoi diagrams (Edelsbrunner, 1987). The next theorem shows some properties of HVD useful for characterizing the optimal bounds $\bar{f}$ and $\underline{f}$.

Theorem 3.
(i) $C^t \neq \emptyset \Leftrightarrow \|\tilde{w}^t - \tilde{w}^\tau\| > \eta^{t\tau}, \forall \tau \neq t$.
(ii) $C^t \cap C^\tau = \emptyset$, $t \neq \tau$.
(iii) $\bigcup_{t=1}^{T} [C^t] = \mathbb{R}^n$, *where* $[C^t]$ *is the closure of* C^t.

Proof. $\Rightarrow$: Suppose that $C^t \neq \emptyset$, then a w exists such that $\|w - \tilde{w}^t\| - \|w - \tilde{w}^\tau\| < \eta^{\tau t}, \forall \tau \neq t$. But the triangular inequality implies: $\|w - \tilde{w}^\tau\| - \|w - \tilde{w}^t\| \leqslant \|\tilde{w}^t - \tilde{w}^\tau\|$, then: $\|\tilde{w}^t - \tilde{w}^\tau\| > -\eta^{\tau t} = \eta^{t\tau}, \forall \tau \neq t$.

$\Leftarrow$: $\|\tilde{w}^t - \tilde{w}^\tau\| > \eta^{t\tau} = -\eta^{\tau t}, \forall \tau \Rightarrow -\|\tilde{w}^t - \tilde{w}^\tau\| < \eta^{\tau t}, \forall \tau \Rightarrow \|\tilde{w}^t - \tilde{w}^t\| - \|\tilde{w}^t - \tilde{w}^\tau\| < \eta^{\tau t}, \forall \tau \Rightarrow \tilde{w}^t \in C^t \Rightarrow C^t \neq \emptyset$.

(ii) $C^t \cap C^\tau = (S^{t1} \cap S^{t2} \cap \cdots) \cap (S^{\tau 1} \cap S^{\tau 2} \cap \cdots) = \cdots \cap S^{t\tau} \cap S^{\tau t} \cap \cdots = \emptyset$, being $S^{t\tau} \cap S^{\tau t} = \emptyset$ by definition.

(iii) Let ξ^t, $t = 1, 2, \dots, T$ a collection of values such that $\eta^{t\tau} = \xi^t - \xi^\tau$, $t, \tau = 1, 2, \dots, T$. The function $g(w) = \min_t(\xi^t + \|w - \tilde{w}^t\|)$, is everywhere defined on $\mathbb{R}^n$ then, for $\forall w \in \mathbb{R}^n$, there exists at least a t such that $\xi^t + \|w - \tilde{w}^t\| \leqslant \xi^\tau + \|w - \tilde{w}^\tau\|$, $\forall \tau$. This implies that $w \in S^{t\tau} \cup H^{t\tau}, \forall \tau$, i.e. $w \in [C^t]$. Thus, every $w \in \mathbb{R}^n$ belongs to at least one set $[C^t]$, then $\bigcup_{t=1}^{\mathrm{T}} [C^t] = \mathbb{R}^n$. □

This result shows that the non-empty cells of a HVD give a complete partition of $\mathbb{R}^n$, so that any $w \in \mathbb{R}^n$ belongs to some $(n-1)$-dimensional hyperbola $H^{t\tau}$ or to one (and only one) cell C^t.

Now, for given $\bar{f}$ and $\underline{f}$, consider the HVD $\bar{V}$ and $\underline{V}$ defined as

$$\bar{V} \doteq V(\tilde{W}^{\mathrm{T}}, \bar{\eta}), \quad \underline{V} \doteq V(\tilde{W}^{\mathrm{T}}, \underline{\eta}),$$

where $\bar{\eta}^{\tau t} = (\bar{h}^\tau - \bar{h}^t)/\gamma$, $\underline{\eta}^{\tau t} = (\underline{h}^t - \underline{h}^\tau)/\gamma$. Let $\bar{C}^t$, $t = 1, 2, \dots, T$ be the cells of $\bar{V}$ and $\underline{C}^t$, $t = 1, 2, \dots, T$ be the cells of $\underline{V}$.

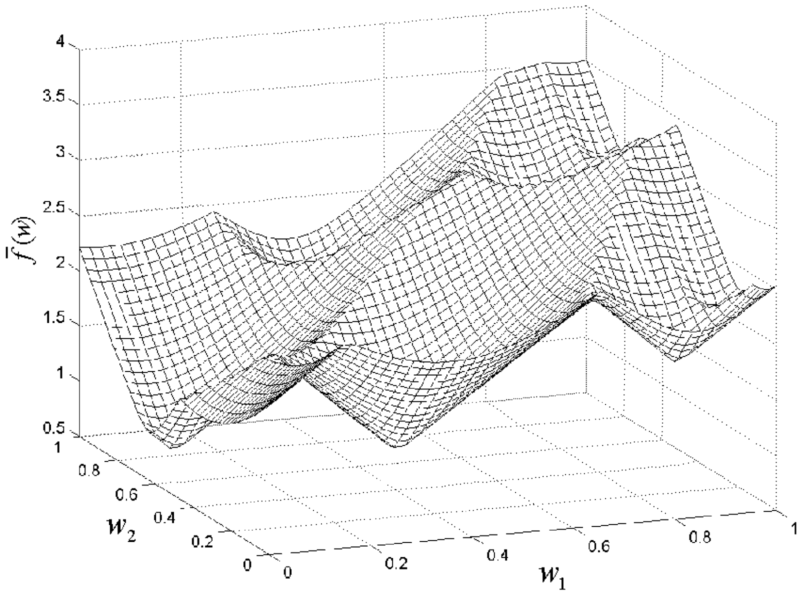

Fig. 2. Example of optimal upper bound $\bar{f}(w)$.

The following result and the comments below show the connection between the HVD $\bar{V}$ and $\underline{V}$ and the optimal bounds $\bar{f}(w)$ and $\underline{f}(w)$.

Theorem 4.

(i) *Let $\bar{C}^t$ be a non-empty cell of $\bar{V}$. Then*: $\bar{f}(w) = \bar{h}^t + \gamma\|w - \tilde{w}^t\|, \forall w \in \bar{C}^t$.

(ii) *Let $\underline{C}^t$ be a non-empty cell of $\underline{V}$. Then*: $\underline{f}(w) = \underline{h}^t - \gamma\|w - \tilde{w}^t\|, \forall w \in \underline{C}^t$.

Proof. From the definition of HVD we have that, if $w \in \bar{C}^t$, then $\|w-\tilde{w}^t\| - \|w-\tilde{w}^\tau\| < \bar{\eta}^{\tau t}, \tau = 1, 2, \ldots, T$. But $\bar{\eta}^{\tau t} = (\bar{h}^\tau - \bar{h}^t)/\gamma$, then $\bar{h}^t + \gamma\|w - \tilde{w}^t\| < \bar{h}^\tau + \gamma\|w - \tilde{w}^\tau\|$, $\tau = 1, 2, \ldots, T$. Since $\bar{f}(w) = \min_{\tau=1,2,\ldots,T}(\bar{h}^\tau + \gamma\|w - \tilde{w}^\tau\|)$, it follows that $\bar{f}(w) = \bar{h}^t + \gamma\|w - \tilde{w}^t\|$, thus proving (i). The proof of (ii) is analogous. □

This theorem shows that, for w belonging to a non-empty cell $\bar{C}^t$, the function $\bar{f}(w)$ is given by the cone in $\mathbb{R}^n \times \mathbb{R}$ defined by the equation $y = \bar{h}^t + \gamma\|w - \tilde{w}^t\|$, with vertex of coordinates $(\tilde{w}^t, \bar{h}^t)$ and axis along the y-dimension. Since from Theorem 3 the non-empty cells of $\bar{V}$ give a complete partition of the regressor space $\mathbb{R}^n$, $\bar{f}$ is a piece-wise conic function over a suitable partition of $\mathbb{R}^n$ that can be derived from the HVD $\bar{V}$. Indeed, the intersection of two cones $y = \bar{h}^t + \gamma\|w - \tilde{w}^t\|$ and $y = \bar{h}^\tau + \gamma\|w - \tilde{w}^\tau\|$, projected on $\mathbb{R}^n$ gives the hyperbola $\bar{H}^{t\tau} = \{w \in \mathbb{R}^n : \|w-\tilde{w}^t\| - \|w-\tilde{w}^\tau\| = \bar{\eta}^{t\tau}, t \neq \tau\}$ that define the HVD $\bar{V}$. Similar considerations hold for the relation between $\underline{f}$ and $\underline{V}$.

In Figs. 2 and 3 the upper bound $\bar{f}$ and the cell partition of $\bar{V}$ are reported for an example, with $w = (w_1, w_2) \in \mathbb{R}^2$. Note that because of the piece-wise conic nature of $\bar{f}$, the level contours of $\bar{f}$ in each cell are circular.

The next result shows that $\underline{f}$ and $\bar{f}$ are differentiable almost everywhere (a.e.) in W, i.e. except a set of zero measure. Let $\bar{V}^d$ and $\underline{V}^d$ be the sets of the d-faces of the HVD $\bar{V}$ and $\underline{V}$, respectively, with $d < n$. Let $co\underline{M}$ and $co\bar{M}$ be the complements in W of the sets $\underline{M} \doteq \bigcup_{d<n} \underline{V}^d \cup \tilde{W}^{\mathrm{T}}$ and $\bar{M} \doteq \bigcup_{d<n} \bar{V}^d \cup \tilde{W}^{\mathrm{T}}$, i.e. $\underline{M} \cup co\underline{M} = W$ and $\bar{M} \cup co\bar{M} = W$. Note that $\underline{M}$ and $\bar{M}$ are sets of zero measure in $\mathbb{R}^n$. In fact, $\tilde{W}^{\mathrm{T}}$ is a set composed by a finite number of points and $\bar{V}^d$, $\underline{V}^d$ are composed by a finite number of d-dimensional sets, with $d < n$.

Theorem 5. *The functions $\underline{f}$ and $\bar{f}$ are Lipschitz-continuous on W. Moreover*:

(i) *$\underline{f}$ is differentiable* $\forall w \in \mathrm{co}\underline{M}$ *and*:

$$\|\underline{f}'(w)\| = \gamma, \ \forall w \in co\underline{M}$$

(ii) *$\bar{f}$ is differentiable* $\forall w \in \mathrm{co}\bar{M}$ *and*:

$$\|\bar{f}'(w)\| = \gamma, \ \forall w \in co\bar{M}.$$

Proof. Let $\hat{w}, w \in W$ and $\bar{t} = \arg\min_t(\bar{h}^t + \gamma\|w - \tilde{w}^t\|)$. From Theorem 2 it follows $\bar{f}(\hat{w}) \leqslant \bar{h}^{\bar{t}} + \gamma\|\hat{w} - \tilde{w}^{\bar{t}}\|$ and $\bar{f}(w) = \bar{h}^{\bar{t}} + \gamma\|w - \tilde{w}^{\bar{t}}\|$. This implies $\bar{f}(\hat{w}) - \bar{f}(w) \leqslant \gamma(\|\hat{w} - \tilde{w}^{\bar{t}}\| - \|w - \tilde{w}^{\bar{t}}\|) \leqslant \gamma\|\hat{w} - w\|$. Similarly it results that $\bar{f}(\hat{w}) - \bar{f}(w) \geqslant -\gamma\|\hat{w} - w\|$, then $|\bar{f}(\hat{w}) - \bar{f}(w)|/\|\hat{w} - w\| \leqslant \gamma$. This holds for $\forall w, \hat{w} \in W$, then $\bar{f}$ is Lipschitz-continuous on W. The Lipschitz-continuity of $\underline{f}$ can be analogously proven.

Let w an arbitrary point of co$\underline{M}$. Thus w belongs to a set $\hat{\underline{C}}^{\underline{t}} \doteq \underline{C}^{\underline{t}} - \tilde{w}^{\underline{t}}$, where the notation $A - B$ indicates the difference between the sets A and B and $\underline{C}^{\underline{t}}$ is a cell of HVD $\underline{V}$ .Being $\hat{\underline{C}}^{\underline{t}}$ an open set and since from Theorem 4 we have $\underline{f}(w) = \underline{h}^{\underline{t}} - \gamma\|w - \tilde{w}^{\underline{t}}\|, \ \forall w \in \hat{\underline{C}}^{\underline{t}}$, it follows that $\underline{f}$ is differentiable

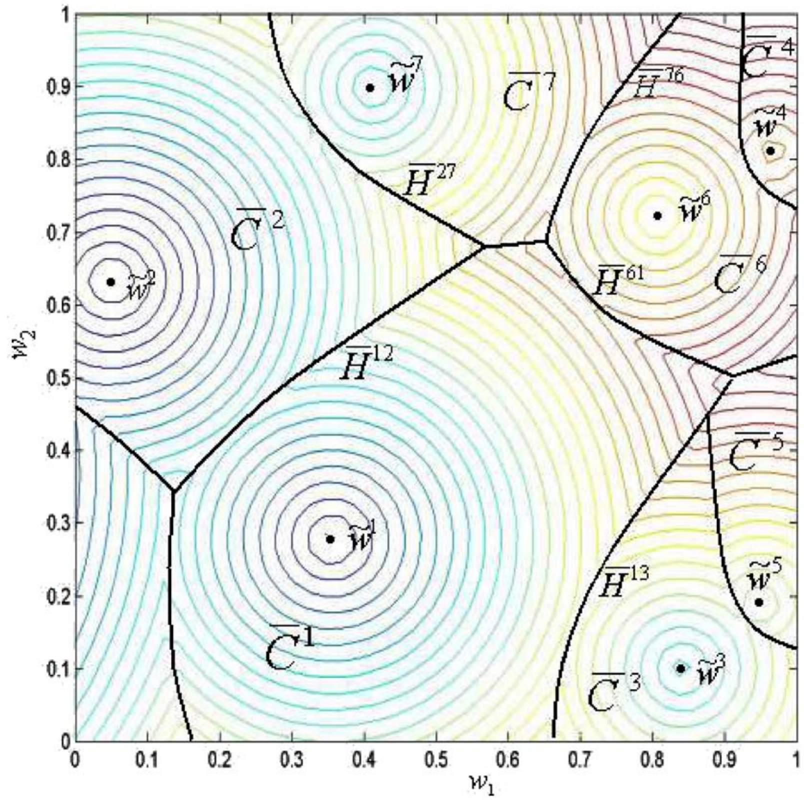

Fig. 3. Level curves of $\bar{f}(w)$ and corresponding HVD $\bar{V}$.

on $\hat{\underline{C}}^{\underline{t}}$ and $\|\underline{f}'(w)\| = \gamma$, thus proving (i). The proof of (ii) is analogous. □

5. Optimal algorithm and estimate

Let the function f_c be defined as

$$f_c(w) \doteq \tfrac{1}{2}[\underline{f}(w) + \bar{f}(w)], \tag{16}$$

where $\underline{f}(w)$ and $\bar{f}(w)$ are given in Theorem 2. We will show that the algorithm $\phi_c(FSS^{\mathrm{T}}) = f_c$ is optimal for any L_p norm. In order to prove this property, we need some preliminary results about $\underline{f}$, $\bar{f}$ and f_c. At first, it is shown that f_c is Lipschitz-continuous and almost everywhere differentiable.

Let $\bar{V}^d$ and $\underline{V}^d$ be the sets of the d-faces of $\bar{V}$ and $\underline{V}$ respectively, with $d < n$. Let coM the complement in W of the set $M \doteq \bigcup_{d<n}(\bar{V}^d \cup \underline{V}^d) \cup \tilde{W}^{\mathrm{T}}$, i.e. $M \cup coM = W$. Note that M is a set of zero measure in $\mathbb{R}^n$. In fact, $\tilde{W}^{\mathrm{T}}$ is a set composed by a finite number of points and $\bar{V}^d, \underline{V}^d$ are sets composed by a finite number of d-dimensional sets, with $d < n$.

Theorem 6.

(i) *The function f_c is Lipschitz-continuous on W.*

(ii) *$f_c(w)$ is differentiable $\forall w \in$ coM and*:

$\|f_c'(w)\| \leqslant \gamma, \ \forall w \in coM$

Proof. (i) In Theorem 5 it has been proven that $\underline{f}$ and $\bar{f}$ are Lipschitz-continuous on W. Then it follows that also f_c is Lipschitz-continuous on W.

(ii) Let w an arbitrary point of co$M$. Thus w belongs to a set $\hat{C}^{\bar{t}\underline{t}} \doteq (\bar{C}^{\bar{t}} \cap \underline{C}^{\underline{t}}) - (\tilde{w}^{\underline{t}} \cup \tilde{w}^{\bar{t}})$, where the notation $A - B$ indicates the difference between the sets A and B, $\bar{C}^{\bar{t}}$ and $\underline{C}^{\underline{t}}$ are cells of HVD $\bar{V}$ and $\underline{V}$, respectively. Being $\hat{C}^{\bar{t}\underline{t}}$ an open set and $f_c(w) = \frac{1}{2}(\underline{h}^{\underline{t}} - \gamma\|w - \tilde{w}^{\underline{t}}\| + \bar{h}^{\bar{t}} + \gamma\|w - \tilde{w}^{\bar{t}}\|)$, $\forall w \in \hat{C}^{\bar{t}\underline{t}}$, it follows that f_c is differentiable on $\hat{C}^{\bar{t}\underline{t}}$. On the other hand, differentiating (16), it follows that $\|f_c'(w)\| \leqslant 1/2[\|\underline{f}'(w)\| + \|\bar{f}'(w)\|]$. This, in view of Theorem 5, implies that $\|f_c'(w)\| \leqslant \gamma$, $\forall w \in coM$. □

We need also the following technical Lemma.

Lemma 1. *Let $\overline{FSS^{\mathrm{T}}}$ be the closure of FSS^{T} with respect to (wrt) norm $\|f\|_S = \|f\|_\infty + \|f'\|_2$. Then, $\underline{f}, \bar{f}$, $f_c \in \overline{FSS^{\mathrm{T}}}$.*

Proof. The linear space $C^1(W)$ of continuously differentiable functions over W, embedded with the norm $\|\cdot\|_S$ is not a Banach space, since it is not complete. Let $F(W)$ be the Banach space obtained by completing $C^1(W)$ wrt $\|\cdot\|_S$. Using the same argument of the proof of theorem 1, $F(W)$ results to be the space of continuous functions, differentiable a.e. in W and

$\overline{FSS^{\rm T}} = \{f \in \overline{\mathscr{F}(\gamma)} : |\tilde{y}^{t+1} - f(\tilde{w}^t)| \leqslant \varepsilon^t,\ t = 1,2,\ldots,T\}$, where $\overline{\mathscr{F}(\gamma)} = \{f \in F(W), \|f'(w)\| \leqslant \gamma$, a.e. in $W\}$.

From Theorem 5 it follows that $\underline{f}, \bar{f} \in \overline{\mathscr{F}(\gamma)}$ and from Theorem 6 that $f_c \in \overline{\mathscr{F}(\gamma)}$. To prove the lemma, it remains to be proved that $|\tilde{y}^{t+1} - f(\tilde{w}^t)| \leqslant \varepsilon^t$, $t = 1,2,\ldots,T$ for $f = \underline{f}, \bar{f}, f_c$.

Since it is assumed that the sufficient condition of Theorem 1 is verified, we have:

$$\bar{f}(\tilde{w}^t) > \tilde{y}^{t+1} - \varepsilon^t, t = 1,2,\ldots,T. \qquad (17)$$

From Theorem 2 it follows:

$$\bar{f}(\tilde{w}^t) \leqslant \tilde{y}^{t+1} + \varepsilon^t, t = 1,2,\ldots,T, \qquad (18)$$

$$\underline{f}(\tilde{w}^t) \geqslant \tilde{y}^{t+1} - \varepsilon^t, t = 1,2,\ldots,T. \qquad (19)$$

Inequalities (17) and (8) are equivalent to the required inequality for $\bar{f}$.

Since $\underline{f}(w) \leqslant \bar{f}(w), \forall w \in W$, from (18) it follows:

$$\underline{f}(\tilde{w}^t) \leqslant \tilde{y}^{t+1} + \varepsilon^t, t = 1,2,\ldots,T. \qquad (20)$$

Inequalities (19) and (20) are equivalent to the required inequality for $\underline{f}$.

From (17) and (19) it follows that:

$$\underline{f}(\tilde{w}^t) + \bar{f}(\tilde{w}^t) \geqslant 2(\tilde{y}^{t+1} - \varepsilon^t), t = 1,2,\ldots,T. \qquad (21)$$

and from (18) and (20) that:

$$\underline{f}(\tilde{w}^t) + \bar{f}(\tilde{w}^t) \leqslant 2(\tilde{y}^{t+1} + \varepsilon^t), t = 1,2,\ldots,T. \qquad (22)$$

Inequalities (21) and (22) are equivalent to the required inequality for f_c, thus concluding the proof of the claim. □

Now we can prove the main result of this section.

Theorem 7. *For any $L_p(W)$ norm, with $p \in [1,\infty]$:*

(i) The identification algorithm $\phi_c(FSS^{\rm T}) = f_c$ is optimal.

(ii) $E(f_c) = \frac{1}{2}\|\bar{f} - \underline{f}\|_p = r_I = \inf_\phi E[\phi(FSS^{\rm T})]$.

Proof. From Theorems 5 and 6 it follows that $\bar{f}, \underline{f}$ and f_c are bounded on W which is bounded. Then, $\bar{f}, \underline{f}, f_c \in L_p(W)$. The diameter of $FSS^{\rm T}$ is

$$d(FSS^{\rm T}) = \sup_{f_1,f_2 \in FSS^{\rm T}} \|f_1 - f_2\|_p \leqslant \left[\int_W |\sup_{f_1 \in FSS^{\rm T}} f_1(w) - \inf_{f_2 \in FSS^{\rm T}} f_2(w)|^p \,\mathrm{d}w\right]^{1/p} = \|\bar{f} - \underline{f}\|_p. \qquad (23)$$

From Lemma 1, since $FSS^{\rm T}$ is dense in $\overline{FSS^{\rm T}}$ wrt $\|\cdot\|_S$, it follows that $\forall \delta > 0, \exists \bar{f}_\delta, \underline{f}_\delta \in FSS^{\rm T}$ such that $\|\bar{f} - \bar{f}_\delta\|_S = \|\bar{f} - \bar{f}_\delta\|_\infty + \|\bar{f}' - \bar{f}'_\delta\|_2 < \delta$, $\|\underline{f} - \underline{f}_\delta\|_S = \|\underline{f} - \underline{f}_\delta\|_\infty + \|\underline{f}' - \underline{f}'_\delta\|_2 < \delta$ and consequently $\|\bar{f} - \bar{f}_\delta\|_\infty < \delta$ and $\|\underline{f} - \underline{f}_\delta\|_\infty < \delta$. Then

$$\|\bar{f}_\delta - \underline{f}_\delta\|_p \geqslant \|\bar{f} - \underline{f}\|_p - \|\bar{f}_\delta - \bar{f}\|_p - \|\underline{f} - \underline{f}_\delta\|_p = \|\bar{f} - \underline{f}\|_p - 2\delta\Omega_p, \qquad (24)$$

where $\Omega_p = [\int_W \,\mathrm{d}w]^{1/p}$. Since δ can be taken as small as desired, (23) and (24) implies that

$$d(FSS^{\rm T}) = \|\bar{f} - \underline{f}\|_p. \qquad (25)$$

On the other hand,

$$E(f_c) = \sup_{f \in FSS^{\rm T}} \|f - f_c\|_p \leqslant \int_W \left[\sup_{f \in FSS^{\rm T}} |f(w) - f_c(w)|^p \,\mathrm{d}w\right]^{1/p} = \frac{1}{2}\|\bar{f} - \underline{f}\|_p = \frac{1}{2} d(FSS^{\rm T}). \qquad (26)$$

From Definitions 3 and 4 we have:

$$r_I \doteq \inf_{\hat{f}} \sup_{f \in FSS^{\rm T}} \|\hat{f} - f\|_p = r(FSS^{\rm T}) \leqslant E(f_c). \qquad (27)$$

Then, from (26), (27) and the well known relations $r(FSS^{\rm T}) \geqslant d(\overline{FSS^{\rm T}})/2$ (Traub & Woźniakowski, 1980), it follows that $E(f_c) = r_I = \frac{1}{2}\|\bar{f} - \underline{f}\|_p$ thus proving claims (i) and (ii). □

Note that the optimal estimate f_c is a Chebicheff center of $FSS^{\rm T}$ in L_p norm for any $p \in [1,\infty]$, i.e.:

$$\sup_{\tilde{f} \in FSS^{\rm T}} \|\tilde{f} - f_c\|_p = \inf_f \sup_{\tilde{f} \in FSS^{\rm T}} \|\tilde{f} - f\|_p$$

but it does not belong to $FSS^{\rm T}$, since it is not differentiable everywhere. However, functions belonging to $FSS^{\rm T}$ approximating f_c in L_p norm with arbitrary precision can be found, as stated in the following result, which is an immediate consequence of Lemma 1.

Theorem 8. *$\forall \delta > 0$, $\forall p \in [1,\infty]$, $\exists f_\delta \in FSS^{\rm T}$ such that $\|f_\delta - f_c\|_p < \delta$.*

Proof. From Theorem 6 in Section 5 it follows that $f_c \in \overline{FSS^{\rm T}}$. Since $\overline{FSS^{\rm T}}$ is the closure of $FSS^{\rm T}$ wrt the norm $\|f\|_S$, $FSS^{\rm T}$ is dense in $\overline{FSS^{\rm T}}$, and consequently $\forall \rho > 0$ a function $f_\rho \in FSS^{\rm T}$ can be found such that $\|f_c - f_\rho\|_S < \rho$. Since $\|f\|_S \geqslant \|f\|_\infty, \forall f$, we have that

$$\|f_c - f_\rho\|_\infty < \rho. \qquad (28)$$

Moreover, since $\|f\|_\infty = \sup_{w \in W} |f(w)|$, it follows that $|f_c(w) - f_\rho(w)| < \rho, \forall w$. This implies:

$$\|f_c - f_\rho\|_p < \rho\Omega_p, \qquad (29)$$

where $\Omega_p = \left[\int_W \,\mathrm{d}w\right]^{1/p}$. The claim follows from (28) and (29), by taking $f_\delta = f_\rho$ with $\rho \leqslant \delta$, $p = \infty$ and $\rho \leqslant \delta/\Omega_p$, $p \in [1,\infty)$. □

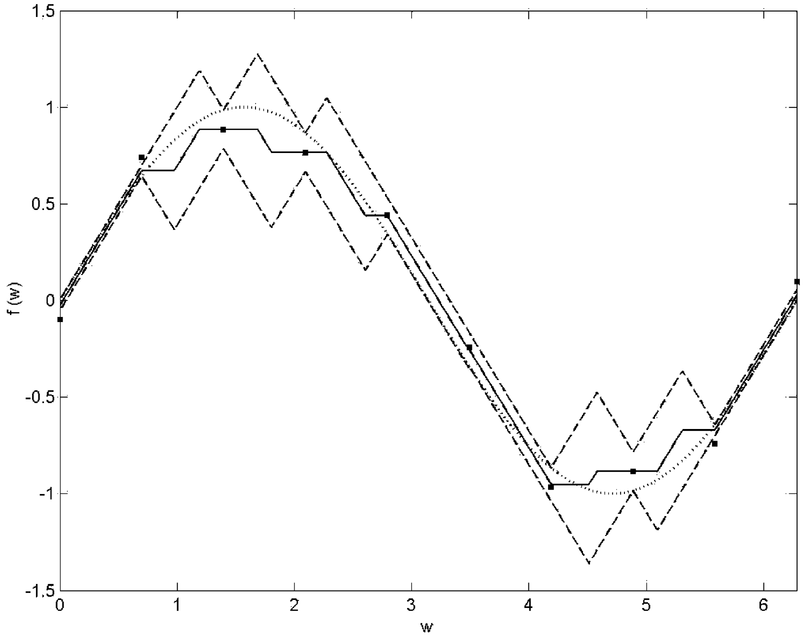

Fig. 4. $f_o(w)$: dotted line. Measurements: squares. $f_c(w)$: solid line. $\bar{f}(w), \underline{f}(w)$: dashed lines.

In the case of one-dimensional regressor space ($n=1$), the functional properties of $f_c(w)$ are quite simple. In particular, $f_c(w)$ is piece-wise linear as shown in Fig. 4, where the regression function $f_o(w)=\sin(w)$ is identified from $T=10$ noise corrupted measurements, assuming $\gamma=1$ and $\varepsilon^t=0.1$, $\forall t$. In the case $n \geqslant 2$, $f_c(w)$ is no more piece-wise linear, but has a more complex behaviour. Indeed, from Theorem 3 it follows that any given w belongs to some cell $\bar{C}^t$ of the HVD $\bar{V}$, related to $\bar{f}$, and to some cell $\underline{C}^\tau$ of the HVD $\underline{V}$, related to $\underline{f}$, i.e. $w \in \bar{C}^t \cap \underline{C}^\tau$. From definition (16) of $f_c(w)$ and from Theorem 4 it follows that $f_c(w)=1/2(\bar{h}^t + \gamma\|w-\tilde{w}^t\| + \underline{h}^\tau - \gamma\|w-\tilde{w}^\tau\|)$, i.e the sum of two cones with vertices coordinates $(\tilde{w}^t, \bar{h}^t)$ and $(\tilde{w}^\tau, \underline{h}^\tau)$. If dimension of regressor space is $n=1$, the sum of two cones is a straight line with angular coefficient 0 and $\pm\gamma$ and then $f_c(w)$ is piece-wise linear. This is not the case if $n \geqslant 2$. Recalling that $\tilde{w}^t \in \bar{C}^t \cap \underline{C}^t$, if this set is not empty, it follows that in such a neighborhood of $\tilde{w}^t$, $f_c(w)$= const$=\tilde{y}^{\,t+1}$. Note that such a neighborhood may be empty, see (i) of Theorem 3. In these cases, $f_c(w) \neq \tilde{y}^{\,t+1}$ for w in a neighborhood of $\tilde{w}^t$. In the example of Fig. 4, this happens for $t=1,2,6,9,10$. For $w \in \bar{C}^t \cap \underline{C}^\tau$ with $t \neq \tau$, $f_c(w)$ is nor a linear variety nor a cone, but a more complex surface.

6. Local assumptions and regressors scaling

6.1. Local assumptions

So far a global bound on $\|f_o'(w)\|$ over all W is assumed. However, a local approach can be taken in order to obtain improvements in identification accuracy, e.g. by assuming different bounds γ_k on suitable partitions W_k of W. This is similar to what done in identification of piece-wise linear model, where partitions W_k are looked for, over which $f_o(w)$ can be considered approximately linear, i.e. $f_o'(w) \simeq$ const., $\forall w \in W_k$, (see e.g. Sontag, 1981; Ferrari-Trecate, Muselli, Liberati, & Morari, 2001). However, finding such partitions may be not an easy task. A very simple alternative approach allowing to use local assumptions on f_o, is based on the evaluation of a function f_a approximating f_o(using any desired method) and on the application of the method described in this paper to the residue function $f_\Delta(w) \doteq f_o(w) - f_a(w)$ using the set of values $\Delta y^{t+1} = \tilde{y}^{\,t+1} - f_a(\tilde{w}^t)$, $t=1,2,\ldots,T$. Then, the estimate:

$$f_c^L(w) = f_a(w) + f_\Delta^c(w) \tag{30}$$

is used, where $f_\Delta^c(w)$ is the central estimate of $f_\Delta(w)$ obtained from data Δy^{t+1}, $t=1,2,\ldots,T$.

Assuming a global bound $\|f_\Delta'(w)\| = \|f_o'(w) - f_a'(w)\| \leqslant \gamma_\Delta$ on the residue function f_Δ implies the local bound $\|f_a'(w)\| - \gamma_\Delta \leqslant \|f_o'(w)\| \leqslant \|f_a'(w)\| + \gamma_\Delta$ for function f_o.

If function f_a is chosen as the PE estimate within a parametric model family $f(w,\theta)$, the present "local" approach allows to investigate the effects of neglected dynamics and of possible trapping in local minima during the parameter estimation phase. In fact, if f_a and $f_c^L(w)$ have comparable identification accuracy, a confirmation is obtained that the chosen model family is sufficiently rich to accurately approximate f_o and that a "good" minimum of $V(\theta, \Phi_T)$, if not the global one, is reached. On the other hand, model $f_c^L(w)$ may give accuracy improvements over model f_a in case the chosen model family is not sufficiently rich and/or minimization of $V(\theta, \Phi_T)$ got stuck in a local minimum.

Even in case that the correction term $f^c_\Delta(w)$ results to be negligible, performing the local SM identification method is useful because allows to derive the following finite samples uncertainty bounds on $f_o(w)$:

$$f_a(w) + \underline{f}_\Delta(w) \leqslant f_o(w) \leqslant f_a(w) + \bar{f}_\Delta(w), \forall w \in W. \tag{31}$$

Note that deriving reliable finite samples results on identification accuracy in the context of PE identification methods is at present a largely open problem.

In Milanese and Novara (2003) the linear regression choice $f_a(w) = \theta w$ is investigated. In particular, conditions are given assuring boundedness of simulation error.

6.2. Regressors scaling

The problem of selecting suitable scalings of regressors is now investigated, in order to adapt to the properties of data. Suitable scaling may turn out to be important when the gradient components have quite different magnitudes, (Stenman, Gustafsson, & Ljung, 1996; Wasilkowski & Woźniakowski, 2001). The problem is posed as follows. Some estimates of the quantities $\mu_i = \max_{w \in W} |\partial f_o(w)/\partial w_i|$, $i = 1, 2, \ldots, n$ can be derived (e.g. from a neural approximation of f_o or directly from data). Here, w_i denotes the i-the component of vector $w \in W \subseteq \mathbb{R}^n$. These estimates support the evidence that:

$$f_o \in \mathscr{F}(\gamma)^\mu_\infty \doteq \{f \in C^1 : \|f'(w)\|^\mu_\infty \leqslant \gamma, \forall w \in W\},$$

where $\|x\|^\mu_\infty \doteq \max_{i=1,\ldots,n} |x_i|\mu_i^{-1}$, $\mu_i > 0$ denotes the weighted ℓ_∞ norm.

Then, $f_o \in \mathscr{F}(\gamma)^\mu_\infty$ could be used as prior assumption on the unknown function f_o. Unfortunately, dealing with such a type of prior appears not easy, and weighted l_2 bounds on the gradient are used, of the form:

$$f_o \in \mathscr{F}(\gamma)^v_2 \doteq \{f \in C^1(W) : \|f'(w)\|^v_2 \leqslant \gamma, \forall w \in W\}$$

where $\|x\|^v_2 \doteq \sqrt{\sum_{i=1}^n v_i x_i^2}$.

Outer approximations $\mathscr{F}(\gamma)^v_2 \supseteq \mathscr{F}(\gamma)^\mu_\infty$ can be looked for, by suitably choosing v. Since $\mathscr{F}(\gamma)^v_2 \supseteq \mathscr{F}(\gamma)^\mu_\infty \Leftrightarrow B^v_2 \supseteq B^\mu_\infty$, where $B^v_2 \doteq \{x \in \mathbb{R}^n : \|x\|^v_2 \leqslant 1\}$, $B^\mu_\infty \doteq \{x \in \mathbb{R}^n : \|x\|^\mu_\infty \leqslant 1\}$, the problem is equivalent to look, in the n-dimensional gradient space, for outer approximations of the weighted ℓ_∞ ball B^μ_∞ with a weighted ℓ_2 ball B^v_2. By taking the ratio of the volumes of the two balls as measure of approximation goodness, minimal volume outer approximation is optimal. The following lemma shows how the optimal solution can be obtained.

Lemma 2. *The optimal* (*minimal volume*) *outer approximation B^v_2 of B^μ_∞ is given by $v_i = (n\mu_i^2)^{-1}$, $i = 1, \ldots, n$.*

Proof. Trivial application of Lagrangian multipliers. □

Let us define the scaled regressors $v_i = 1/\sqrt{v_i} w_i$, $i = 1, 2, \ldots, n$ and, with a slight abuse of notation, denote $f(w) = f(w_i, \ldots, w_i) = f(\sqrt{v_1} v_i, \ldots, \sqrt{v_n} v_i) = f(v)$. Then $\partial f(v)/\partial v_i = \partial f(w)/\partial w_i \sqrt{v_i}$ and $\|f'(w)\|^v_2 = \sqrt{\sum_{i=1}^n v_i(\partial f(w)/\partial w_i)^2} = \|f'(v)\|$. Thus, considering the scaled regressors v, a bound on the euclidean norm of gradient is obtained. The results presented in the previous sections can be directly applied by substituting regressors w with scaled regressors v.

7. Summary of the NSM identification process

The main steps of proposed method are now summarized. The case of global assumptions on $\|f'(w)\|$ is considered. Minor modification are required for the case of local assumptions.

(1) Partition the data to be used for the identification in two parts. The first T data, called estimation data, are used in steps 2,3,4 and 6. The remaining data, called calibration data, are used in step 5 for the selection of $\gamma, \varepsilon_r, \varepsilon_a$ values.

Define the range of interest of regressors:

$$w \in W = \{[\underline{w}_1, \bar{w}_1] \times \cdots \times [\underline{w}_n, \bar{w}_n]\}.$$

(2) Perform a preliminary rough estimate $f_b(w)$ of $f_o(w)$ by some identification method.

(3) Compute $\mu_i = \max_{w \in W} |\partial f_b(w)/\partial w_i|$, $i = 1, 2, \ldots, n$ and consider the scaled regressors:

$$v_i = \frac{w_i}{\sqrt{v_i}}, \quad i = 1, 2, \ldots, n, \quad v_i = (n\mu_i^2)^{-1}.$$

Let

$$V = \{[\underline{w}_1/\sqrt{v_1}, \bar{w}_1/\sqrt{v_1}] \times \cdots \times [\underline{w}_n/\sqrt{v_n}, \bar{w}_n/\sqrt{v_n}]\}$$

and $f(v) = f(\sqrt{v_1} v_i, \ldots, \sqrt{v_n} v_i)$.

(4) Compute the surface $\gamma^*(\varepsilon_r, \varepsilon_a)$ defined by (14) on a suitable range of values of $(\varepsilon_r, \varepsilon_a)$. This task is performed by means of theorem 1, using not the original regressors w^t but the scaled regressors v^t instead.

(5) Select $(\gamma, \varepsilon_r, \varepsilon_a)$ values in the validated region. A reasonable choice is $\hat{\gamma} \cong \max_{v \in V} \|f_b'(w)\|$, $\hat{\varepsilon}_a \cong$ accuracy of device used for y^t measurements and $\hat{\varepsilon}_r$ in the validated region, giving the minimum of $RMSE(\varepsilon_r, \hat{\gamma}, \hat{\varepsilon}_a)$, where $RMSE(\varepsilon_r, \gamma, \varepsilon_a)$ is the simulation error on the calibration set of the regression model computed as in step 6 for given $\gamma, \varepsilon_r, \varepsilon_a$.

(6) The identified regression model is

$$y^{t+1} = f_c(v^t) = f_c\left(\frac{y^t}{\sqrt{v_1}}, \ldots, \frac{y^{t-n_y+1}}{\sqrt{v_{n_y+1}}}, \ldots, \frac{u_m^{t-n_m+1}}{\sqrt{v_n}}\right),$$

where $f_c(v) = \frac{1}{2}[\underline{f}(v) + \bar{f}(v)]$ and $\underline{f}(v)$, $\bar{f}(v)$ are given in Theorem 2 by substituting w with scaled regressors v, and using the selected values $\hat{\gamma}, \hat{\varepsilon}_r, \hat{\varepsilon}_a$.

It must be remarked that in the paper and in the above described procedure, given values of regression orders n_y, $n_1, \dots, n_m$ are considered. In practical applications, these values are seldom known and have to be suitably chosen. Several approaches have been proposed in the literature for this task (Sjöberg et al., 1995). In the examples presented in Section 8, we used the simple and widely used approach of performing the identification for different choices of regression orders, evaluating for each identified model an index of its predictive ability and choosing the regression orders giving the best index. In the presented examples, the index is $RMSE(\hat{\varepsilon}_r, \hat{\gamma}, \hat{\varepsilon}_a)$.

8. Examples

8.1. Example 1: water heater

In this example we investigate the water heater identification problem considered also in (Stenman et al., 1996). The system (see Fig. 5) is constituted by a volume of water heated by a resistor element. The heating process can be described by an output variable, i.e. the temperature T^t of the water, and by an input variable, i.e. the voltage u^t that controls the resistor by means of a thyristor. It is expected that the main nonlinearity is due to nonlinear characteristic of the thyristor.

The dataset is composed by a series of 3000 samples of T^t and u^t recorded every 3 s. According to what done in Stenman et al. (1996), the dataset is divided into an identification set, composed by the first 2000 data, and a validation set, composed by the remaining 1000 data (see Fig. 6). The identification set was used to identify two Nonlinear Set Membership models and a neural networks model. The validation set was used to test the identified models in simulation and to compare the simulation performances with those presented in Stenman et al. (1996), where a just in time model (JIT) and a fuzzy model are considered.

The following regression has been considered in all these methods:

$$y^{t+1} = f(w^t) \quad w^t = [T^t \; T^{t-1} \; u^{t-3} u^{t-4}]. \tag{32}$$

This is the choice of regressors made in Stenman et al. (1996) and no other sets of regressors have been looked for, in order to allow a fair comparison with the results reported in Stenman et al. (1996).

8.1.1. Neural network model NN

The NN model is obtained by taking:

$$f(w^t) = \psi(w^t),$$

where the function ψ is a one hidden layer neural network (see e.g. Hertz, Krogh, & Palmer, 1991; Vapnik, 1995) composed by r neurons:

$$\psi(w^t) = \sum_{l=1}^{r} \alpha_l \sigma(\beta_l w^t - \lambda_l) + \zeta. \tag{33}$$

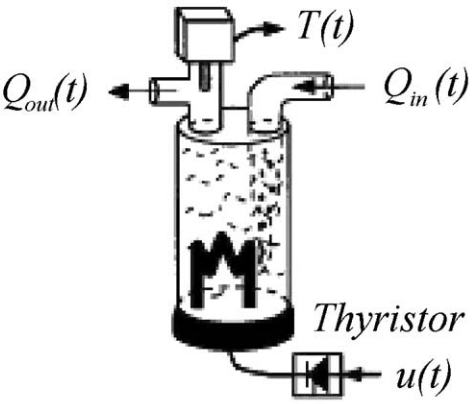

Fig. 5. Example 1. Heater.

Here $\alpha_l, \lambda_l, \zeta \in \mathbb{R}$, $\beta_l \in \mathbb{R}^n$, are parameters and $\sigma(x) = 2/(1 + e^{-2x}) - 1$ is a sigmoidal function. Several neural networks of the form (33) with different values of r (from $r=3$ to $r=20$) have been trained on the identification set. A neural network with $r = 8$, showing the best performances in simulation, has been chosen for the model NN.

8.1.2. Nonlinear set membership model NSMG

The NSMG model has been obtained assuming a global bound on the function gradient. According to Step 1 of Section 7, the identification dataset has been splitted in two subset. The data from 1 to 1700 have been taken as estimation set, the data from 1701 to 2000 have been taken as calibration set. The range of interest of regressors has been chosen as $W = \{[10, 50] \times [10, 50] \times [0, 1] \times [0, 1]\}$. The estimate $f_b(w) = \psi(w)$ has been used at Step 2, where $\psi(w)$ is the neural network used for the model NN.

The scaled regressors v^t used for the estimation of central function are

$$v^t = \left[\frac{T^t}{\sqrt{v_1}} \; \frac{T^{t-1}}{\sqrt{v_2}} \; \frac{u^{t-3}}{\sqrt{v_3}} \; \frac{u^{t-4}}{\sqrt{v_4}} \right], \tag{34}$$

where the scaling vector $v = [2.4 \; 6.9 \; 0.01 \; 0.01]$ has been derived as described in Step 3.

A global bound $\|f'(v)\| \leqslant \gamma$ and a noise bound $|d^t| \leqslant \varepsilon_r |y^{t+1}| + \varepsilon_a$, $\forall t$, have been assumed.

According to Step 5, $\hat{\gamma}=2.7$ and $\hat{\varepsilon}_a=0.5$ have been chosen. Indeed, it resulted $\max_{v \in V} \|f_b'(v)\| = 2.5$ and the accuracy of temperature sensor is ±0.5°C. In order to choose ε_r, the validation surface $\gamma^*(\varepsilon_r, \varepsilon_a)$ has been computed according to Step 4. The section $\gamma^*(\varepsilon_r, 0.5)$ is shown in Fig. 7, where the level curves of $RMSE(\varepsilon_r, \gamma, 0.5)$ are also reported. The value of $\hat{\varepsilon}_r = 0.03$ has been chosen, which corresponds to the minimum of $RMSE(\varepsilon_r, 2.7, 0.5)$ in the validated region. Note anyway that the choice of these values is not critical, as shown in Fig. 7, where $RMSE(\varepsilon_r, \gamma)$ is not very variable in a quite large neighborhood of the selected values $(\hat{\varepsilon}_r, \hat{\gamma}) = (0.03, 2.7)$.

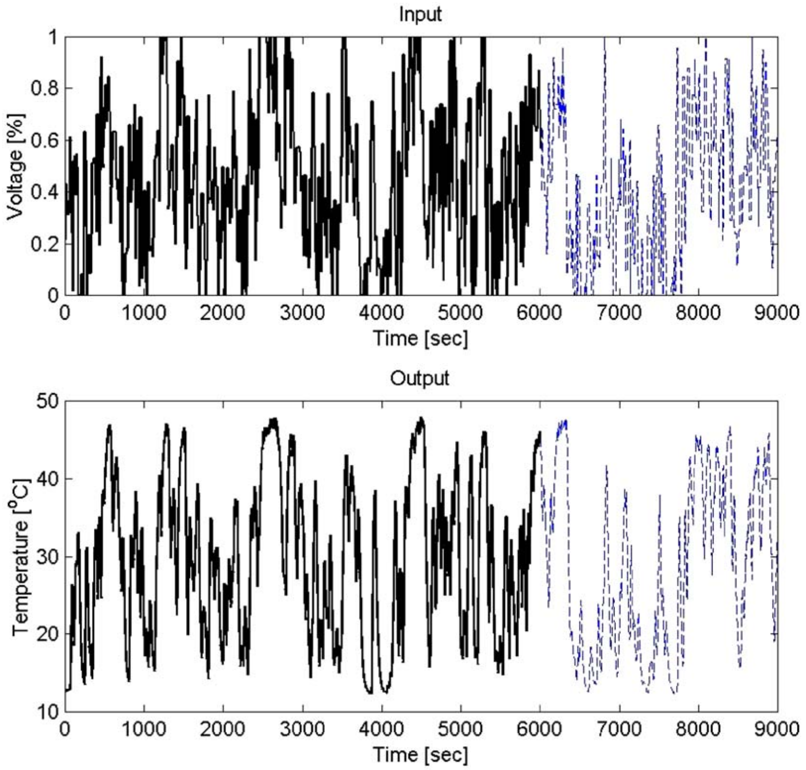

Fig. 6. Example 1. Heater dataset. Bold line: identification set. Dashed line: testing set.

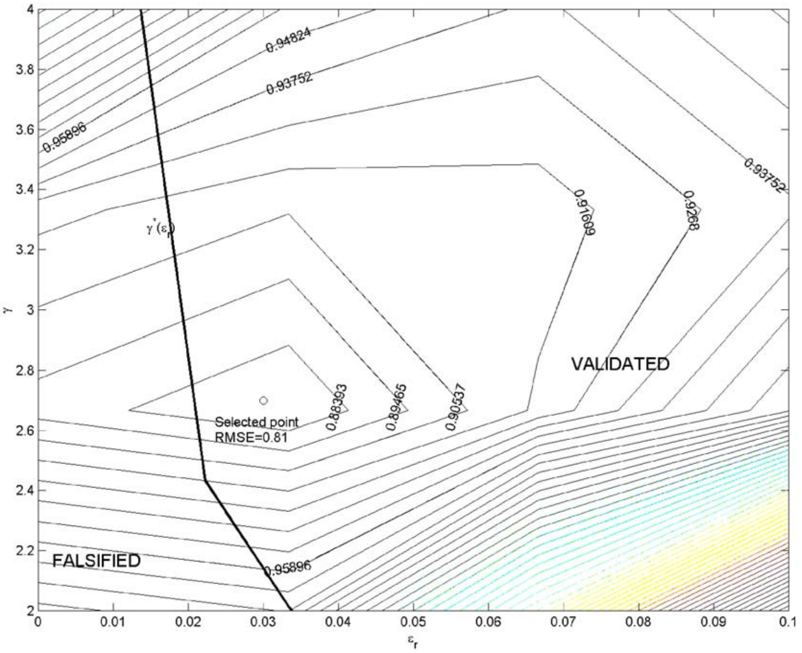

Fig. 7. Example 1.$RMSE(\varepsilon_r, \gamma)$ level curves (thin lines) and validation curve $\gamma^*(\varepsilon_r, 0.5)$ (bold line) for NSMG model.

The NSMG model is the regression model $y^{t+1} = f_c(v^t)$, where f_c is the central function derived as described in Step 6.

8.1.3. Nonlinear set membership model NSML

The NSML model was obtained by means of the local approach described in Section 6.1 by taking $f_a(w) = \psi(w)$,

Table 1
Example 1: simulation errors on the validation dataset

Model	NSMG	NSML	NN	JIT	Fuzzy
RMSE	0.96	0.70	0.80	0.89	1.02

where $\psi(w)$ is the neural network used for model NN. The procedure of Section 7 has been applied considering the residue function $f_\Delta(w) \doteq f_o(w) - f_a(w)$ and using the data $\Delta y^{t+1} = \tilde{y}^{\,t+1} - \psi(\tilde{w}^t)$, $t = 1, 2, \ldots, T$. The same splitting of identification data and same range of regressors as for the model NSMG are used.

The estimate $f_b(w) = \psi_\Delta(w)$ has been used at step 2, where $\psi_\Delta(w)$ is a neural sigmoidal network with 6 neurons estimated using the residue data Δy^t.

The regressors have been scaled as in (34) by means of the scaling vector $v = [6.8\ 6.7\ 0.12\ 0.12]$, derived according to Step 3.

A bound $\|f'_\Delta(v)\| \leqslant \gamma_\Delta$ on the gradient of residue function $f_\Delta(v) = f_o(v) - f_a(v)$ and a noise bound $|d^t| \leqslant \varepsilon_r |\Delta y^{t+1}| + \varepsilon_a$, $\forall t$, have been assumed.

According to Step 5, $\hat{\gamma}_\Delta = 0.07$ and $\hat{\varepsilon}_a = 0.5$ have been chosen. Indeed, it resulted $\max_{v \in V} \|f'_b(v)\| = 0.06$ and the accuracy of temperature sensor is ±0.5°C. In order to choose ε_r, the validation surface $\gamma^*(\varepsilon_r, \varepsilon_a)$ has been computed according to Step 4. The value of $\hat{\varepsilon}_r = 0.85$ has been chosen. The NSML model is the regression model $y^{t+1} = f^L_c(v^t) = \psi(v^t) + f^c_\Delta(v^t)$, where f^c_Δ is the central estimate of f_Δ derived as described in Step 6.

The models NN, NSMG and NSML have been tested in simulation of the validation dataset. In Table 1 the root mean squared errors (RMSE) obtained on such validation data not used for identification are reported and compared with those obtained by the JIT and Fuzzy models considered in Stenman et al. (1996). In Fig. 8 simulation of model NSML on the validation data is shown.

8.2. Example 2: Mechanical system with input saturation

A set of 6000 data has been generated from the following nonlinear system $y^{t+1} = 1.8 y^t - 0.82 y^{t-1} + 0.0024 \sin(y^{t-1}) + 0.047 \tanh(3u^t)$, representing a discrete-time approximation of a mass-spring-damper system with linear spring, nonlinear damper and static nonlinearity on the input (see Fig. 9). Input u is the force acting on the mass and output y is the mass position.

A random input of amplitude $\leqslant 1$ has been used. The output data have been corrupted by a uniform random additive noise of amplitude $\leqslant 0.025$. The first 5000 data have been used for model identification, the remaining 1000 data, called validation set, have been used for model testing.

The following models have been identified. For all of them, regressions of the form

$$y^{t+1} = f(w^t), \quad w^t = [y^t\ y^{t-1}\ u^t]$$

have been considered.

8.2.1. Nonlinear set membership local model NSML

The NSML model was obtained by means of the local approach described in Section 6.1 by taking $f_a(w) = \theta w$, where $\theta = [1.8\ {-0.81}\ 0.06]$ has been estimated by means of the Matlab Systems Identification Toolbox using the output error estimation method. The procedure of Section 7 has been applied considering the residue function $f_\Delta(w) \doteq f_o(w) - \theta w$ and using the data $\Delta y^{t+1} = \tilde{y}^{\,t+1} - \theta(\tilde{w}^t)$, $t = 1, 2, \ldots, T$. The first 4000 data of the identification set have been taken as estimation set and the last 1000 data have been used as calibration set. No regressor scaling has been performed. A bound $\|f'_\Delta(v)\| \leqslant \gamma_\Delta$ on the gradient of residue function $f_\Delta(w) = f_o(w) - \theta w$ and an absolute noise bound $|d^t| \leqslant \varepsilon_a$, $\forall t$ have been assumed.

According to Step 5, $\hat{\gamma}_\Delta = 0.0024$ and $\hat{\varepsilon}_a = 0.08$ have been chosen. The NSML model is the regression model $y^{t+1} = f^L_c(w^t) = \theta w^t + f^c_\Delta(w^t)$, where f^c_Δ is the central estimate of f_Δ, derived as described in Step 6.

8.2.2. Neural network models NARX and NOE

The NARX and NOE models have been obtained considering one hidden layer neural networks of the form (33) for the regression function. Several NARX and NOE neural network models with different values of r (from $r = 3$ to 16) have been trained on the estimation set using the Matlab Neural Networks Toolbox. The NARX model with $r = 8$ has been chosen, showing the best simulation performances. All the NOE identified models got stuck on (possibly) local minima during the training phase, providing bad simulation performances. The best result has been obtained by using as starting point the parameters of the selected NARX model.

In Table 2 the root mean square errors obtained by the identified models on the validation dataset are reported. The simulation error is indicated as $\mathrm{RMSE_S}$ and the one-step ahead prediction error is indicated as $\mathrm{RMSE_P}$. It can be noted that the accuracy improvements of the NSML model over the NARX and NOE models, though moderate for one-step ahead prediction, are quite significant in simulation.

In Fig. 10, a portion of validation data and NSML and NOE models simulation are shown.

8.3. Example 3: Vehicles with controlled suspensions

Identification of vehicle vertical dynamics is considered in this example. Models of vehicles vertical dynamics are very important tools in the automotive field, especially in view of the increasing diffusion of controlled suspension systems (Krtolica & Hrovat, 1992; Lu & DePoyster, 2002). Indeed, accurate models may allow efficient tuning of control algorithms in computer simulation environment, thus significantly reducing the expensive in-vehicle tuning effort.

Identification is performed on simulated data obtained by the half-car model with controlled suspensions shown in Fig. 11. Identification results using experimental data are reported in Milanese, Novara, Mastronardi, and Amoroso

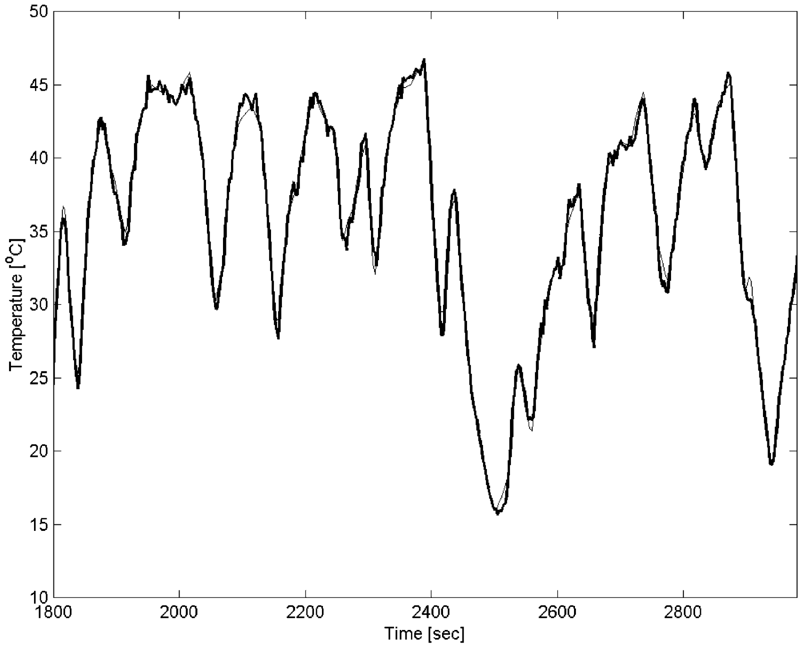

Fig. 8. Example 1. Simulation of model NSML (thin line) on the validation data (bold line).

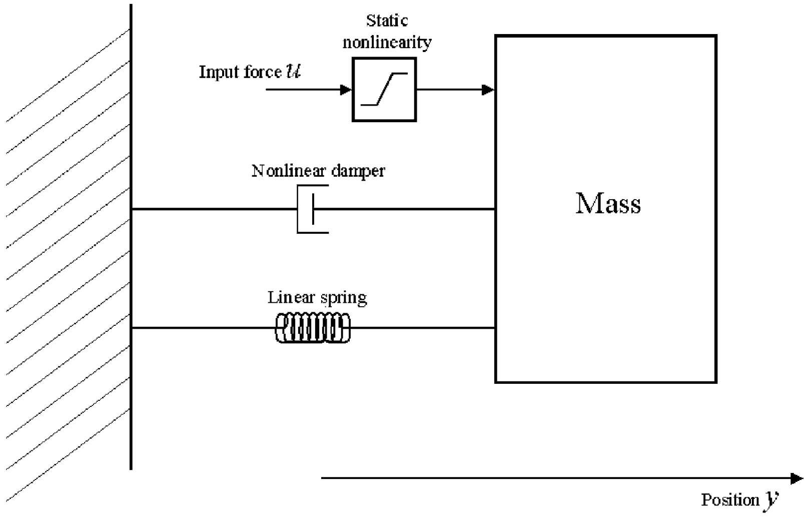

Fig. 9. Example 2. Nonlinear mass-spring-damper system.

Table 2
Example 2. One-step ahead prediction and simulation errors

Model	NSM	NN_{narx}	NN_{noe}
$RMSE_P$	0.005	0.008	0.009
$RMSE_S$	0.091	0.299	0.262

(2004b) and Milanese, Novara, Gabrielli, and Tenneriello (2004a).

The half-car model, called for short “true system”, has been implemented in Simulink in order to obtain data simulating a possible experimental setup, characterized by type of exciting input, experiment length, variables to be measured and accuracy of sensors. The vehicle is assumed to travel in a constant speed $V = 60$ km/h. The main variables describing the model are:

- p_{rf} and p_{rr}: f ront and rear road profiles.
- i_{sf} and i_{sr}: control currents of front and rear suspensions.
- a_{cf} and a_{cr}: front and rear chassis vertical accelerations.
- p_{cf} and p_{cr}: front and rear chassis vertical positions.
- p_{wf} and p_{wr}: front and rear wheels vertical positions.

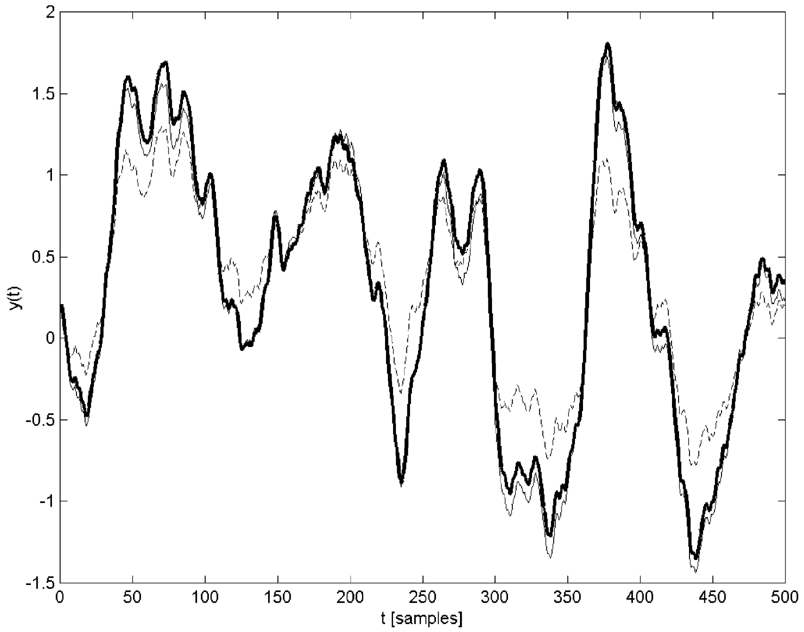

Fig. 10. Example 2. Validation set: data (bold line), NSML simulation (thin line) and NOE simulation (dashed line).

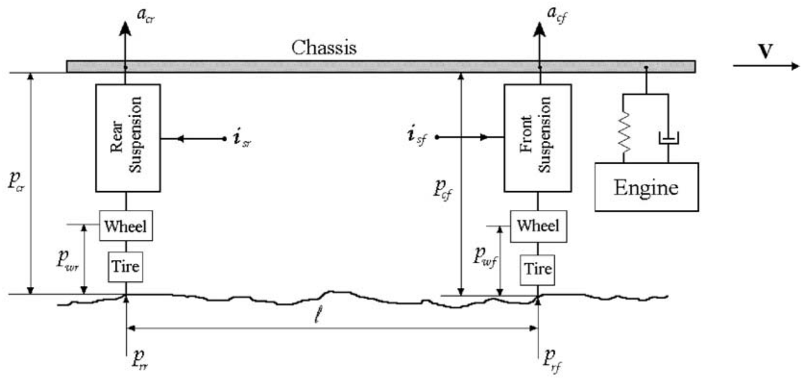

Fig. 11. Example 3. The half-car model.

It is considered that the road profile $p_{\mathrm{rf}}(t)$ is known, that $p_{\mathrm{rr}}(t) = p_{\mathrm{rf}}(t - \ell/V)$, that currents $i_{\mathrm{sf}}(t)$ and $i_{\mathrm{sr}}(t)$ can be measured with a precision of 3.75% and that variables $a_{\mathrm{cf}}(t)$, $a_{\mathrm{cr}}(t)$ can be measured with a precision of 5%.

The chassis, the engine and the wheels are simulated as rigid bodies. The following static nonlinear characteristic has been assumed for tires:

$$F_1(t) = F_{1E}(\Delta p_1(t)) + \beta_1 \Delta v_1(t),$$

where F_1 is the tire force, Δp_1 and Δv_1 are the differences of position and velocity at the extremes of tire, $\beta_1 = 10\,000$ Ns/m and $F_{1E}(\Delta p_1)$ is shown in Fig. 12b. The following nonlinear characteristic has been assumed for controlled suspensions:

$$F_2(t) = K_2 \Delta p_2(t) + F_{2D}(\Delta v_2(t), i(t)),$$

where F_2 is the suspension force, Δp_2 and Δv_2 are the differences of position and velocity at the extremes of suspension, i is the control current, $K_2 = 17200$ N/m, $F_{2D}(\Delta v_2, i)$ is shown in Fig. 12a for the two extreme values $i = 0\ A$ and $1.6A$.

A dataset has been generated from "true system" simulation, for a period of 24 s, using a random profile with amplitude $\leqslant 4$ cm. The dataset consists of the values of p_{rf}, p_{rr}, i_{sf}, i_{sr}, a_{cf} recorded with a sampling time of $\tau = \frac{1}{512}$ s. The sequence of each measured variables is composed of 12 280 samples. The values of a_{cf} have been corrupted by uniformly distributed noises of relative amplitude 5% and the values of i_{sf} and i_{sr} have been corrupted by uniformly distributed noises of relative amplitude 3.75%. The dataset related to the first 20 s, called identification dataset, has been used for models identification. The dataset related to the last 4 s, called validation dataset, has been used to test the

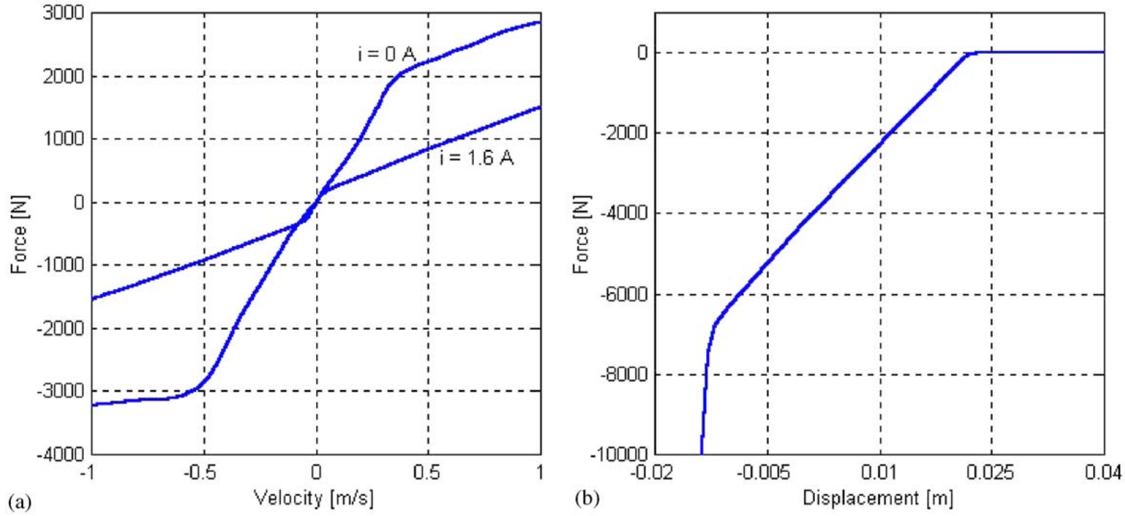

Fig. 12. Example 3. (a) Force–velocity chracteristic $F_{2\mathrm{D}}$ of suspension. (b) Force–displacement characteristic $F_{1\mathrm{E}}$ of tires.

simulation accuracy of identified models. The experimental setup simulated here has been chosen because not too complex to be realized in actual experiment on real car (Milanese et al., 2004a, b). Two models, relating front chassis accelerations to the road profile at the sampling times, have been identified from the identification dataset. The models are of the form

$$y^{t+1} = f(w^t),$$

$$w^t = [y^t \dots y^{t-7}\ u_1^t \dots u_1^{t-2} u_2^t \dots u_2^{t-2}\ u_3^t\ u_4^t]$$

with $y^t = a_{\mathrm{cf}}(t\tau)$, $u_1^t = p_{\mathrm{rf}}(t\tau)$, $u_2^t = p_{\mathrm{rr}}(t\tau)$, $u_3^t = i_{\mathrm{sf}}(t\tau)$ and $u_4^t = i_{\mathrm{sr}}(t\tau)$. The regressors orders $n_y = 8$, $n_1 = 3$, $n_2 = 3$, $n_3 = 1$, $n_4 = 1$ have been chosen as described in Section 7.

8.3.1. Neural network model NN

The NN model is obtained by taking:

$$f(w^t) = \psi(w^t),$$

where the function ψ is a one hidden layer sigmoidal neural network of the form (33). Several neural networks with different values of r (from $r = 3$ to 20) have been trained on the identification set. A neural network with $r = 6$, showing the best performances in simulation, has been chosen for the model NN.

8.3.2. Nonlinear Set Membership model NSML

The NSML model has been obtained by means of the local approach described in Section 6.1 with $f_a(w) = \psi(w)$, where $\psi(w)$ is the neural network used for model NN. The procedure of Section 7 has been applied considering the residue function $f_\Delta(w) \doteq f_o(w) - f_a(w)$ and using the residue data $\Delta y^{t+1} = \tilde{y}^{t+1} - \psi(\tilde{w}^t)$, $t = 1, 2, \dots, 12\,280$. The 7680 data corresponding to the first 15 s of the identification set have been taken as estimation set, the 1536 data corresponding to the last 5 s have been taken as calibration set.

Table 3
Example 3: simulation errors on the validation dataset

Model	NSML	NN
RMSE	0.65	0.66

The scaling vector $v = [0.95, 0.95, 0.95, 0.95, 0.95, 0.95, 0.95, 0.95, 10, 10, 10, 1, 1, 1, 1, 0.5]$ has been evaluated according to step 3 with $f_b(w) = \psi_\Delta(w)$, where $\psi_\Delta(w)$ is a neural sigmoidal network with 6 neurons, estimated using the residue data Δy^t.

A bound $\|f'_\Delta(v)\| \leqslant \gamma_\Delta$ on the gradient of residue function $f_\Delta(v) = f_o(v) - f_a(v)$ and a noise bound $|d^t| \leqslant \varepsilon_r |\Delta y^{t+1}| + \varepsilon_a$, $\forall t$, have been assumed. The values $\hat{\varepsilon}_a = 0.05$, $\hat{\varepsilon}_r = 0.7$, $\hat{\gamma} = 0.2$ have been chosen, according to the procedure described in Steps 4 and 5 and using the residue data Δy^t.

The NSML model is the regression model $y^{t+1} = f_L^c(v^t) = \psi(v^t) + f_\Delta^c(v^t)$, where f_Δ^c is the central function derived as described in Step 6.

The models NN and NSML have been tested in simulation on the validation set. The root mean square simulation errors (RMSE) obtained by models NN and NSML on this dataset are reported in Table 3. In Fig. 13, a portion of "true" data and of the ones simulated by the identified NSML model are reported.

9. Conclusions

In the paper, a method for identification of nonlinear systems described in the form of nonlinear regressions has been presented, based on a SM approach. The novelty is that the method does not assume to know the functional form of nonlinear regression function, in contrast with most

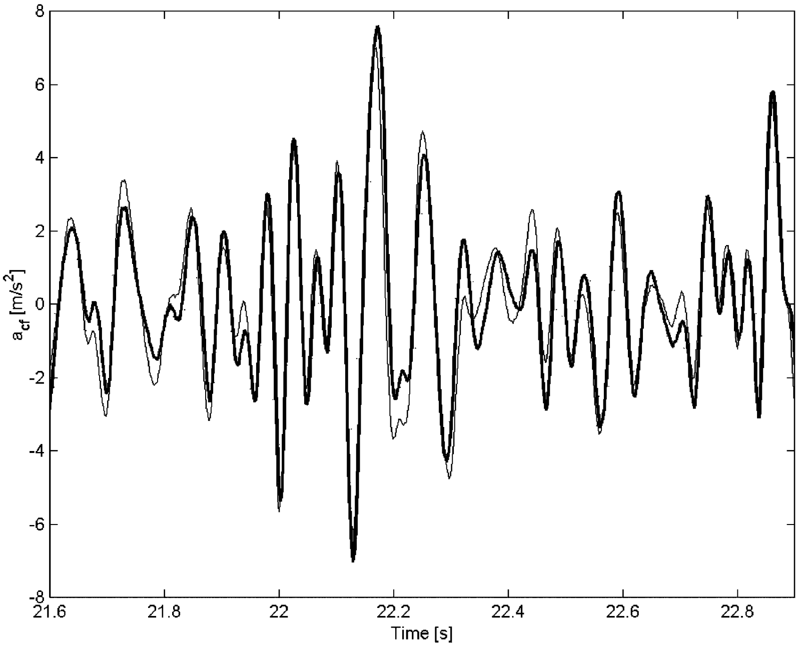

Fig. 13. Example 3. Front chassis accelerations: “true” (bold line), NSML model simulation (thin line).

methods of the literature, which assume that it belongs to a finitely parametrized family. Thus, the method does not require extensive searches of such functional form and reduces the effects of modeling errors due to the use of approximate forms. Moreover, the noise is assumed only to be bounded, in contrast with standard approaches, relying on statistical assumptions such as stationarity, uncorrelation, etc., whose validity is difficult to be reliably checked and anyway is lost in presence of approximate modeling. On the basis of these theoretical features, it is expected that models obtained by means of the proposed method may have good performance and exhibit good robustness versus imprecise knowledge of involved nonlinearities and of noise properties. These expectations appear to be confirmed by the examples presented here and by the applications reported in Novara and Milanese (2001), Milanese and Novara (2002), Milanese et al. (2004a) and Milanese, Novara, Volta, and Finzi (2003).

It can be noted that the present SM approach and the methods based on estimation within a parametric model family can be usefully applied in a complementary way. Indeed, the NSML models, obtained from local assumptions on the gradient of f_o as described in Section 6, simply consists in a correction term to an initial estimate of the regression function, applying the basic SM method to the residuals. Supposing that the initial regression is obtained by a parametric estimation within a parametric model family, if the correction term is negligible, a confirmation is obtained that the chosen model family is sufficiently rich to accurately approximate f_o and that a “good” minimum of the loss function, if not the global one, is reached. On the other hand, NSML model may give accuracy improvements over the estimated parametric model in case the chosen model family is not sufficiently rich and/or minimization of got stuck in a local minimum. In any case, the SM identification method allows to derive the finite samples uncertainty bounds on $f_o(w)$, which can be useful for further robustness investigations, such as guaranteed stability of simulation errors (Milanese & Novara, 2003) or robust control design.

In conclusion, the new approach to nonlinear systems identification presented in this paper appears to be quite promising and at present is under test on larger classes of applications. Also, many important problems remain open, such as complexity analysis, input excitation conditions, stability of solutions of identified models, and are currently under investigation.

Acknowledgements

The authors thank Professors Lennart Ljung and Anders Stenman for having provided the heater dataset. The authors also thank Ing. Luca Pivano for its contribution to development of Example 3.

References

Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transaction on Information Theory*, *39*, 930–945.

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

Edelsbrunner, H. (1987). *Algorithms in combinatorial geometry*. Berlin: Springer.

Ferrari-Trecate, G., Muselli, M., Liberati, D., & Morari, M. (2001). A clustering technique for the identification of piecewise affine systems. In A. Sangiovanni-Vincentelli, & M.D. Di Benedetto (Eds.), *Hybrid systems: computation and control, Lecture notes in computer science*. Berlin: Springer.

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

Hertz, J., Krogh, A., & Palmer, G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison-Wesley, Santa Fe Institute studies in the sciences of complexity.

Hornik, K., Stinchcombe, M., White, H., & Auer, P. (1994). Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, *6*, 1262–1275.

Krtolica, R., & Hrovat, D. (1992). Optimal active suspension control based on a half-car model: An analytical solution. *IEEE Transaction on Automatic Control 37*(4), 528–532.

Lu, J., & DePoyster, M. (2002). Multiobjective optimal suspension control to achieve integrated ride and handling performance. *IEEE Transactions on Control Systems Technology*, *10*(6), 807–821.

Milanese, M., Norton, J., Piet Lahanier, H., & Walter, E. (1996). *Bounding approaches to system identification*. New York: Plenum Press.

Milanese, M., & Novara, C. (2002). Nonlinear Set Membership prediction of river flow. In *Proceedings of the 41st IEEE conference on decision and control*, Las Vegas, Nevada.

Milanese, M., & Novara, C. (2003). Model quality in nonlinear SM identification. In *Proceedings of the 42nd IEEE conference on decision and control*, Maui, Hawaii.

Milanese, M., Novara, C., Gabrielli, P., & Tenneriello, L. (2004a). Experimental modeling of controlled suspension vehicles from onboard sensors. In *1st IFAC Symposium on Advances in Automotive Control*, Salerno, Italy.

Milanese, M., Novara, C., Mastronardi, F., & Amoroso, D. (2004b). Experimental modeling of vertical dynamics of vehicles with controlled suspensions. In *SAE world congress*, Detroit, Michigan.

Milanese, M., Novara, C., Volta, M., & Finzi, G. (2003). Prediction models for air quality management in urban areas. In *CIRA annual meeting*, Modena, Italy.

Milanese, M., & Tempo, R. (1985). Optimal algorithms theory for robust estimation and prediction. *IEEE Transaction on Automatic Control*, *30*, 730–738.

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

Novak, E. (1988). *Deterministic and stochastic error bounds in numerical analysis*, Vol. 1349. Berlin: Springer.

Novara, C., & Milanese, M. (2001). Set Membership prediction of nonlinear time series. In *Proceedings of the 40th IEEE conference on decision and control*, Orlando, FL (pp. 2131–2136).

Partington, J. R. (1997). *Interpolation, identification and sampling*, Vol. 17. Oxford, New York: Clarendon Press.

Popper, K. R. (1969). *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Rontedge and Kegan Paul.

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

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

Sontag, E. D. (1981). Nonlinear regulation. The piecewise linear approach. *IEEE Transactions on Automatic Control*, *26*, 346–357.

Sontag, E. D. (1992). Neural nets as systems models and controllers. In *Seventh Yale workshop on adaptive and learning systems* (pp. 73–79).

Stenman, A., Gustafsson, F., & Ljung, . (1996). Just in time models for dynamical systems. In *Proceedings of the 35th IEEE conference on decision and control*, Kobe, Japan (pp. 1115–1120).

Traub, J. F., Wasilkowski, G. W., & Woźniakowski, H. (1988). *Information-based complexity*. New York: Academic Press.

Traub, J. F., & Woźniakowski, H. (1980). *A general theory of optimal algorithms*. New York: Academic Press.

Vapnik, V. (1995). *The nature of statistical learning theory*. Berlin: Springer.

Wasilkowski, G. W., & Woźniakowski, H. (2001). Complexity of weighted approximation over R^d. *Journal of Complexity*, *17*, 722–740.

Mario Milanese graduated in Electronic Engineering at Politecnico di Torino in 1967. From 1968 to 1972 he was Teaching Assistant at Politecnico di Torino, from 1972 to 1980 Associate Professor of System Theory at Università di Torino. From 1980 he is Full Professor of System Theory at Politecnico di Torino. From 1982 to 1987 he was head of the Dipartimento di Automatica e Informatica at Politecnnico di Torino. His research interests include robust identification, prediction and control of uncertain systems, and applications to biomedical, automotive, aerospace, financial and environmental problems. He is author of more than 180 papers in international journals and conference proceedings. He is editor of the books "Robustness in Identification and Control", Plenum Press, 1989 and "Bounding Approaches to System Identification", Plenum Press, 1996.

Carlo Novara was born in Imperia, Italy in 1970. He received the Laurea degree in Physics from the Facoltà di Scienze M.F.N., Università di Torino in 1996 and the Ph.D. degree in Information and System Engineering from Politecnico di Torino in 2002. He held a visiting position at the Department of Mechanical Engineering, University of California at Berkeley in 2001. He holds currently a post-doc position at the Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino, Italy. His research interests include nonlinear systems identification, robust identification, time series prediction and automotive and environmental applications.