Set Membership prediction of nonlinear time series

Mario Milanese and Carlo Novara[1]

Abstract. In this paper a prediction method for nonlinear time series based on a Set Membership (SM) approach is proposed. The method does not require the choice of the functional form of the model used for prediction, but assumes a bound on the rate of variation of the regression function defining the model. At the contrary, most of the existing prediction methods need the choice of a functional form of the regression function or of state equations (piecewise linear, quadratic, etc.) and this choice is usually the result of heuristic searches. These searches may be quite time consuming, and lead only to approximate model structures, whose errors may be responsible of bad propagation of prediction errors, especially for the multi-step ahead prediction. Moreover, the method proposed in this paper assumes only that the noise is bounded, in contrast with statistical approaches, which rely on noise assumptions such as stationarity, ergodicity, uncorrelation, type of distribution, etc. The validity of these assumptions may be difficult to be reliably tested in many applications and is certainly lost in presence of approximate modeling. In the present SM approach, using a result developed in [1], the values of the bounds on the gradient of the regression function and on the noise can be suitably assessed to verify the validity tests. Two almost optimal prediction algorithms are then derived, the second one having improved optimal properties over the first one, at the expense of an increased computational complexity. The method is tested and compared with other literature methods on the well known Wolf Sunspot Numbers series, widely used in the time series literature as a benchmark test, and on the prediction of vertical dynamics of vehicles with controlled suspensions. A simulation example is also presented to investigate how much conservative the SM approach may be in the most adverse situation where data are generated by a linear AR model driven by i.i.d. gaussian white noise and the SM prediction is compared with the optimal statistical predictor, which makes use of the exact assumptions.

1 Introduction

Time series prediction is a fundamental problem in most fields of science and technology ranging from signal processing [2, 3], to coding [4], finance and economy [5, 6], hydrometeorology [7, 8, 9], weather and climate [10], chemistry [11], production and distribution of goods and services [12]. Prediction plays also an important role in control [13]. A huge literature is available, presenting various approaches and methodology for its solution, essentially based on the identification of some model of the mechanism generating the data and on the use of the identified model for prediction (see e.g. [11, 14, 15, 16]).

Consider a nonlinear dynamic system of the form:

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

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

[1] M. Milanese and C. Novara are with the Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino, Italy. E-mails: `mario.milanese@polito.it, carlo.novara@polito.it` .

A set of noise corrupted measurements $\widetilde{y}^t$ and $\widetilde{w}^t$ of y^t and w^t, $t = 1, 2, ..., T$ is available and the aim is to derive a prediction $\widehat{y}^{T+k}$ of y^{T+k}, possibly giving "small" prediction error $|\widehat{y}^{T+k} - y^{T+k}|$.

Since data are finite and noise corrupted, providing only limited information on f_o, whatever prediction $\widehat{y}^{T+k}$ is used, no finite bound on the prediction error can be derived if no information is available on f_o and on noise. Indeed, it is well known that determining a model from a finite set of data without any prior knowledge about the system is an ill-posed problem, in the sense that a unique model may not exist, or it may not depend continuously on data [17]. The information on f_o is typically given by considering that it belongs to a finitely parametrized set of functions $K \doteq \{f(p), \, p \in \Re^q\}$. Measured data are then used to derive an estimate $\widehat{p}$ of free parameters p and $f(\widehat{p})$ is used to make predictions. In some cases, the knowledge of the laws governing the system (mechanical, economical, biological, 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 black-box parametrizations. Basic to this approach is the proper choice of the set of functions $f(p)$, typically realized by some search on different functional forms, starting from the simplest ones, such as linear models, possibly after nonlinear transformations of data (logarithmic, square root, etc.) and moving to more complex ones, such as piecewise linear, bilinear, neural networks, etc. [18, 19, 20]. This search may be quite time consuming, and in any case leads to approximate model structures only. Evaluating the effects of such approximation on the propagation of the prediction error appear to be a formidable problem, since most of the properties of prediction methods are derived under the assumption that f_o belongs to the chosen family of functions $f(p)$, [11, 14].

In this paper we propose an alternative approach based on a Set Membership (SM) framework which proved useful in linear systems identification with approximate models [21, 22, 23, 24, 25]. The approach has connection with Information Based Complexity (IBC) methods for evaluating functionals of multivariable functions with bounded derivatives, from the knowledge of a finite number of their values (see e.g. [26, 27, 28] and the references therein). In the IBC literature, noise free measurements are typically assumed, and weaker optimality concepts are considered than the one of the present paper. The proposed method does not assume to know the functional form of f_o, but uses only some information on its regularity, given by bounds on the gradients of f_o. In this way, the problem of considering approximate functional forms of f_o is circumvented. Moreover it is assumed that measurements are corrupted by bounded noise, in contrast with statistical approaches, which rely on assumptions such as stationarity, ergodicity, uncorrelation, type of distribution, etc. The validity of these assumptions may be difficult to be reliably tested in many applications and is certainly lost in presence of approximate modeling, when $f_o \notin K \doteq \{f(p), p \in \Re^q\}$. The same assumptions of the present paper has been used in [1] for the identification problem of finding an estimate $\widehat{f}$ of f_o, giving "small" L_p identification error $||f_o - \widehat{f}||_p$. As a matter of fact, the two problems are related, in the sense that they are two specific instances of the general problem of making an inference on the unknown system, described by the operator $I(f^o, W_T)$, where $W_T \doteq [w_1, w_2, ..., w_T]$. Specifically, $I(f^o, W_T) = f^o$ if the desired inference is the identification of the unknown function f^o, as considered in [1], and $I(f^o, W_T) = f^o(w_T)$ if the inference is one-step prediction, as considered in the present paper. However, the two problems (identification and prediction) are different problems with different solutions. In particular, the optimal solution of the identification problem, derived in [1] does not provide optimal prediction. Indeed, in the prediction problem considered in the present paper, finding an optimal algorithm appears to be difficult, as it often happens in SM-IBC contexts [23, 26]. However, two almost optimal Nonlinear Set Membership (NSM) prediction algorithms are derived, the second one having improved optimal properties over the first one, at the expense of increased computational complexity.

The paper is organized as follows. In section 2, the prediction problem is formulated in the SM framework, defining system and noise assumptions, prediction error, validation and optimality concepts. In section 3, two almost optimal algorithms and upper and lower bounds on their worst case prediction error are derived. Conditions are also given, under which such algorithms give actually optimal prediction and their error bound are actually the minimal worst case prediction error achievable by any algorithm. In section 4, the approach is tested and compared with other literature methods on the well known Wolf Sunspot Numbers series, widely used in the time series literature as a benchmark test, and on the prediction of vertical dynamics of vehicles with controlled suspensions. A simulation example is also presented to investigate how much conservative the SM approach may be in the most adverse situation where data are generated by a linear AR model driven by i.i.d. gaussian white noise and the NSM prediction algorithms are compared with the optimal statistical predictor, which makes use of the exact assumptions.

2 Nonlinear Set Membership prediction and optimality properties

In this section, the prediction problem is formulated in a Set Membership framework, see e.g. [21, 22, 23]. Consider that a set of noise corrupted data $\widetilde{Y}^T = [\widetilde{y}^1, \widetilde{y}^2, ..., \widetilde{y}^T]$ and $\widetilde{W}^T = [\widetilde{w}^1, \widetilde{w}^2, ..., \widetilde{w}^T]$ generated by (1) is available. Then:

$$\widetilde{y}^{t+1} = f_o(\widetilde{w}^t) + d^t, \; t = 1, 2, .., T-1$$

where the term d^t accounts for the fact y^{t+1} and w^t are not exactly known. The aim is to obtain an estimate $\widehat{y}^{T+k}$ of y^{T+k}, possibly giving small $k-$step ahead prediction error $|\widehat{y}^{T+k} - y^{T+k}|$.

It must be noted that no finite bound on the prediction error can be guaranteed, unless some assumptions are made on the function f_o and the noise d. The typical approach in the literature is to assume a given functional form for f_o (linear, bilinear, etc.) and statistical models on the noise sequence. 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 $D^T = [d^1, d^2, ..., d^T]$ is only supposed to be bounded.

Prior assumptions on f_o:

$$f_o \in K \doteq \left\{ f \in C^1(W) : \|f'(w)\| \leq \gamma, \; \forall w \in W \right\}$$

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

■

Prior assumptions on noise:

$$D^T \in \mathcal{D} \quad \doteq \left\{ [d^1, d^2, ..., d^T] : |d^t| \leq \varepsilon^t + \gamma\delta^t,\ t = 1, 2, ..., T \right\}$$

where ε^t and δ^t are the bounds on noises affecting y and w according to $\left|y^{t+1} - \widetilde{y}^{t+1}\right| \leq \varepsilon^t$, $\|w^t - \widetilde{w}^t\| \leq \delta^t$, $t = 1, 2, ..., T$. The rational for this assumption is that $d^t = \widetilde{y}^{t+1} - y^{t+1} + f_o(w^t) - f_o(\widetilde{w}^t)$ and then $|d^t| \leq \left|\widetilde{y}^{t+1} - y^{t+1}\right| + \gamma \|w^t - \widetilde{w}^t\|$.

■

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 *The Feasible Systems Set FSS^T is:*

$$FSS^T \doteq \{f \in K : \left|\widetilde{y}^{t+1} - f\left(\widetilde{w}^t\right)\right| \leq \varepsilon^t + \gamma\delta^t,\ t = 1, 2, ..., T-1\}$$

■

The Feasible Systems Set FSS^T summarizes all the information (measured data and prior information on f_o and noise d) that is available up to time T on the mechanism generating the data. As required in any identification theory, the problem of checking the validity of prior assumptions arises. Indeed, the only thing that can be actually done is to check if prior assumptions are invalidated by data, evaluating if no unfalsified system exists, i.e. if FSS^T is empty. However, it is usual to introduce the concept of prior assumption validation as follows:

Definition 2 *Prior assumptions are considered validated if $FSS^T \neq \emptyset$.*

■

Necessary and sufficient conditions for checking the assumptions validity, are given below in theorem 1, which is reported from [1].

Let us introduce the following quantities:

$$\begin{aligned} \overline{f}\left(w\right) &\doteq \min_{t=1,...,T-1} \left(\overline{h}^t + \gamma \|w - \widetilde{w}^t\|\right) \\ \underline{f}\left(w\right) &\doteq \max_{t=1,...,T-1} \left(\underline{h}^t - \gamma \|w - \widetilde{w}^t\|\right) \end{aligned} \tag{2}$$

where $\overline{h}^t \doteq \widetilde{y}^{t+1} + \varepsilon^t + \gamma\delta^t$ and $\underline{h}^t \doteq \widetilde{y}^{t+1} - \varepsilon^t - \gamma\delta^t$.

Theorem 1 *[1]*

i) A necessary condition for prior assumptions to be validated is:

$$\overline{f}\left(\widetilde{w}^t\right) \geq \underline{h}^t \qquad t = 1, 2, ..., T-1$$

ii) A sufficient condition for prior assumptions to be validated is:

$$\overline{f}\left(\widetilde{w}^t\right) > \underline{h}^t \qquad t = 1, 2, ..., T-1$$

■

Theorem 1 can be used for assessing the values of ε^t, δ^t and γ, in order to have a non-empty FSS^T. In order to not require too much detailed information on bounds ε^t, δ^t, absolute or relative error models for these bounds can be adopted, i.e. $\varepsilon^t = \varepsilon \; \forall t$, $\delta^t = \delta \; \forall t$ or $\varepsilon^t = \varepsilon \left| \widetilde{y}^{t+1} \right| \; \forall t$, $\delta^t = \delta \left\| \widetilde{w}^t \right\| \; \forall t$.

In the space $(\varepsilon, \delta, \gamma)$, the function:

$$\gamma^* (\varepsilon, \delta) \doteq \inf_{FSS^T \neq \emptyset} \gamma \tag{3}$$

represents a surface that separate falsified values of ε, δ and γ from validated ones. Clearly, ε, δ and γ must be chosen in the validated parameters region (see section 7 of [1] and the end of the next section for some more details on the selection of these constants).

Let us now define the error of given prediction algorithm. For the sake of simplicity of exposition, let us focus on one-step ahead prediction. A one-step ahead prediction algorithm ϕ is a function mapping all available information until time T about data, function f_o and noise d, summarized by FSS^T, into the predicted value of y^{T+1}:

$$\widehat{y}^{T+1} = \phi \left(FSS^T \right)$$

Its prediction error $PE = |\widehat{y}^{T+1} - y^{T+1}| = |\phi \left(FSS^T \right) - f_o \left(w^T \right) |$ is not exactly known, since it is only known that $f_o \in FSS^T$ and $w^T \in B_\delta \left(\widetilde{w}^T \right) \doteq \{ w : ||w - \widetilde{w}^T|| \leq \delta^T \}$. Thus, the worst case prediction error defined as:

$$WPE(\, \widehat{y}^{T+1}) \doteq \sup_{f \in FSS^T} \; \sup_{w^T \in B_\delta (\widetilde{w}^T)} | \, \widehat{y}^{T+1} - f \left(w^T \right) |$$

can be used as a measure of prediction accuracy.

Looking for prediction algorithms that minimize the worst case prediction error, leads to the following optimality concepts.

Definition 3 *A prediction $\widehat{y}_o^{T+1}$ is said optimal if:*

$$WPE(\, \widehat{y}_o^{T+1}) = \inf_\phi WPE \left[\phi \left(FSS^T \right) \right]$$

A prediction algorithm ϕ^o is called optimal if:

$$WPE \left[\phi^o \left(FSS^T \right) \right] = \inf_\phi WPE \left[\phi \left(FSS^T \right) \right], \; \forall FSS^T$$

■

Thus, an optimal prediction algorithm gives optimal predictions for any available information up to time T. As it often happens in SM-IBC theory (see e.g.[22, 23, 26]), finding optimal algorithms is in general hard, motivating the interest of deriving simpler algorithms, at the expense of some degradation in the prediction error with respect to an optimal algorithm. In particular, algorithms guaranteeing a degradation in the prediction error of at most 2 are widely considered in the literature, and called "almost optimal", according to the following definition.

Definition 4 *: A prediction algorithm ϕ^{ao} is called almost optimal if:*

$$WPE \left[\phi^{ao} \left(FSS^T \right) \right] \leq 2 \inf_\phi WPE \left[\phi \left(FSS^T \right) \right], \; \forall FSS^T$$

■

3 Almost optimal prediction algorithms

In the identification problem investigated in [1], it has been shown that the function $f_c(w) \doteq \frac{1}{2}\left[\underline{f}(w) + \overline{f}(w)\right]$ is an optimal estimate of f_o, in the sense that it minimizes the worst case L_p norm $||f_o - f_c||_p$, $\forall p \geq 1$.

In analogy to the statical setting, where optimal identification and prediction are strictly related, it can be expected that prediction $f_c(\widetilde{w}^T)$ is optimal. Indeed, it will be shown that such a prediction is in general only almost optimal and that optimality is reached under some conditions.

In order to formulate the results, we need to recall the notion of Hyperbolic Voronoi Diagrams (HVD), a generalization of the standard Voronoi Diagrams, introduced in [1]. Consider the set of points:

$$\widetilde{W}^T \doteq [\widetilde{w}^1, \widetilde{w}^2, ..., \widetilde{w}^T]$$

and a $T \times T$ antisymmetric matrix η. Then define:

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

$$H^{t\tau} \doteq \{w \in \Re^n : \left\|w - \widetilde{w}^t\right\| - \left\|w - \widetilde{w}^\tau\right\| = \eta^{t\tau},\ t \neq \tau\}$$

- The n-dimensional region $S^{t\tau}$ containing $\widetilde{w}^t$:

$$S^{t\tau} \doteq \{w \in \Re^n : \left\|w - \widetilde{w}^t\right\| - \left\|w - \widetilde{w}^\tau\right\| < \eta^{\tau t},\ t \neq \tau\}$$

- The hyperbolic cell C^t:

$$C^t \doteq \bigcap_{\tau \neq t} S^{t\tau}$$

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

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

■

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 [29].

Now, for given $\overline{f}$ and $\underline{f}$ defined in (2), consider the HVD $\overline{V}$ and $\underline{V}$ defined as:

$$\begin{aligned} \overline{V} &\doteq V\left(\widetilde{W}^T, \overline{\eta}\right) \\ \underline{V} &\doteq V\left(\widetilde{W}^T, \underline{\eta}\right) \end{aligned}$$

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

From Theorem 4 of [1] it follows that, for w belonging to a non-empty cell $\overline{C}^t$, the function $\overline{f}(w)$ is given by the cone in $\Re^n \times \Re$ defined by the equation $y = \overline{h}^t + \gamma \left\|w - \widetilde{w}^t\right\|$, with vertex of coordinates $\left(\widetilde{w}^t, \overline{h}^t\right)$ and axis along the y-dimension. Since the non-empty cells of $\overline{V}$ give a complete partition of the regressor space $\Re^n$, $\overline{f}$ is a piece-wise conic function over a suitable partition of $\Re^n$ that can be derived from the HVD $\overline{V}$. Indeed, the intersection of two cones $y = \overline{h}^t + \gamma \|w - \widetilde{w}^t\|$ and $y = \overline{h}^\tau + \gamma \|w - \widetilde{w}^\tau\|$, projected on $\Re^n$ gives the hyperbola $\overline{H}^{t\tau} = \{w \in \Re^n : \|w - \widetilde{w}^t\| - \|w - \widetilde{w}^\tau\| = \overline{\eta}^{t\tau},\ t \neq \tau\}$ that define the HVD $\overline{V}$. Similar considerations hold for the relation between $\underline{f}$ and $\underline{V}$.

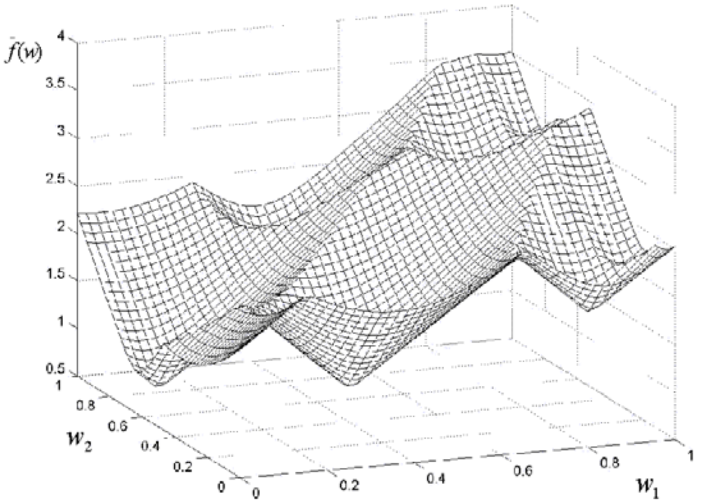

Figure 1: Example of optimal upper bound $\overline{f}(w)$.

In Figures 1 and 2 the function $\overline{f}$ and the cell partition of $\overline{V}$ are reported for an example, with $w = (w_1, w_2) \in \Re^2$. Note that because of the piece-wise conic nature of $\overline{f}$, the level contours of $\overline{f}$ in each cell are circular.

We are now in the position of formulating the main results of the paper, related to the derivation of two almost optimal algorithms and upper and lower bounds on their worst case prediction error. Conditions are also given, under which such algorithms give actually optimal predictions and their error bound are actually the minimal worst case prediction errors achievable by any algorithm.

Since from Theorem 3 of [1], it results that the non-empty cells of a HVD give a complete partition of $\Re^n$, it follows that $\widetilde{w}^T$ belongs to one specific cell of the HVD $\overline{V}$ and to one specific cell of the HVD $\underline{V}$, i.e. $\exists \overline{t}, \underline{t}$ such that $\widetilde{w}^T \in \left[\overline{C}^{\overline{t}}\right]$ and $\widetilde{w}^T \in \left[\underline{C}^{\underline{t}}\right]$, where $\overline{C}^{\overline{t}}$ and $\underline{C}^{\underline{t}}$ are hyperbolic cells of HVD $\overline{V}$ and $\underline{V}$, respectively. Let us define the following quantities:

$$\overline{w}^T \doteq \widetilde{w}^T + \delta^T \frac{\widetilde{w}^T - \widetilde{w}_{\overline{t}}}{||\widetilde{w}^T - \widetilde{w}_{\overline{t}}||} \qquad \underline{w}^T \doteq \widetilde{w}^T + \delta^T \frac{\widetilde{w}^T - \widetilde{w}_{\underline{t}}}{||\widetilde{w}^T - \widetilde{w}_{\underline{t}}||} \tag{4}$$
$$\overline{f}_u\left(\widetilde{w}^T\right) \doteq \overline{f}\left(\widetilde{w}^T\right) + \gamma\delta^T \qquad \underline{f}_l\left(\widetilde{w}^T\right) \doteq \underline{f}\left(\widetilde{w}^T\right) - \gamma\delta^T$$

which are needed in the proof of theorem 2 and in the formulation of theorem 3.

Theorem 2

i) The prediction algorithm:

$$\phi_1^c\left(FSS^T\right) = \frac{1}{2}\left[\underline{f}\left(\widetilde{w}^T\right) + \overline{f}\left(\widetilde{w}^T\right)\right]$$

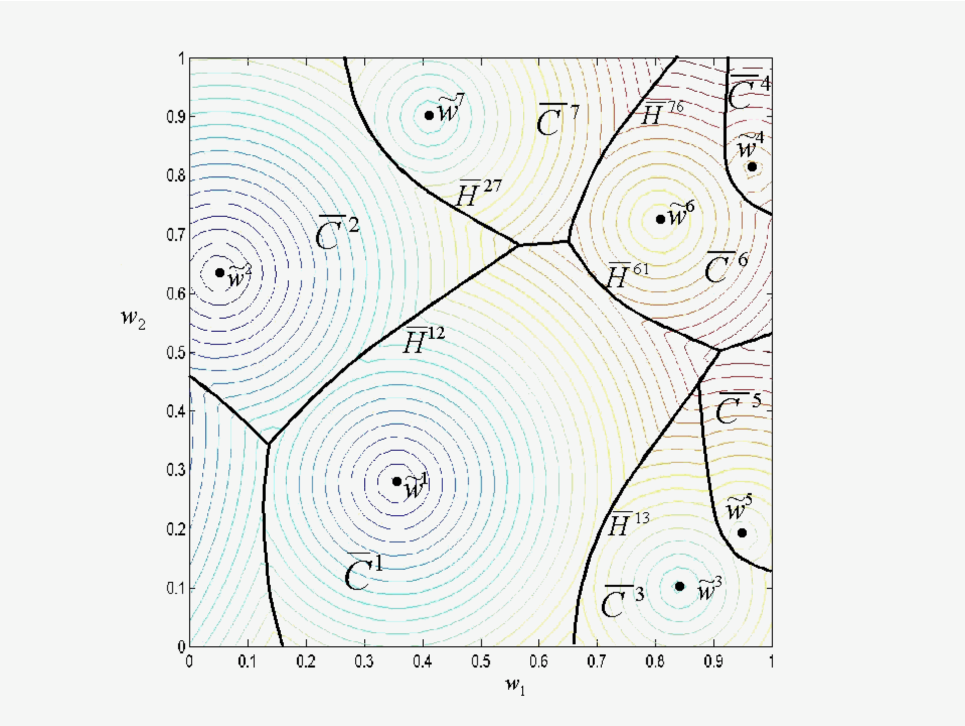

Figure 2: Level curves of $\overline{f}(w)$ and corresponding HVD $\overline{V}$.

is almost optimal, with prediction error bounded as:

$$WPE\left[\phi_1^c\left(FSS^T\right)\right] \leq \frac{1}{2}\left[\overline{f}\left(\widetilde{w}^T\right) - \underline{f}\left(\widetilde{w}^T\right)\right] + \gamma\delta^T$$

ii) If $B_\delta\left(\widetilde{w}^T\right) \subseteq \overline{C}^{\overline{t}} \cap \underline{C}^{\underline{t}}$*, then the prediction* $y_c^{T+1} = \phi_1^c\left(FSS^T\right)$ *is optimal with minimal worst case prediction error given by:*

$$WPE\left[\phi_1^c\left(FSS^T\right)\right] = \frac{1}{2}\left[\overline{f}\left(\widetilde{w}^T\right) - \underline{f}\left(\widetilde{w}^T\right)\right] + \gamma\delta^T$$

Proof.

Let us define the following functions:

$$\begin{array}{l} \underline{f}^*\left(\widetilde{w}^T\right) \doteq \inf_{w \in B_\delta\left(\widetilde{w}^T\right)} \inf_{f \in FSS^T} f\left(w\right) \\ \overline{f}^*\left(\widetilde{w}^T\right) \doteq \sup_{w \in B_\delta\left(\widetilde{w}^T\right)} \sup_{f \in FSS^T} f\left(w\right) \end{array} \qquad (5)$$

Clearly, $\underline{f}^*\left(\widetilde{w}^T\right)$ and $\overline{f}^*\left(\widetilde{w}^T\right)$ are the tightest lower and upper bound of $y^{T+1} = f_o\left(w^T\right)$:

$$\underline{f}^*\left(\widetilde{w}^T\right) \leq f_o\left(w^T\right) \leq \overline{f}^*\left(\widetilde{w}^T\right)$$

As well known from Set Membership theory (see e.g. [21, 22, 23]), the prediction:

$$y_o^{T+1} = \frac{1}{2}\left[\underline{f}^*\left(\widetilde{w}^T\right) + \overline{f}^*\left(\widetilde{w}^T\right)\right] \doteq \phi^o\left(FSS^T\right) \qquad (6)$$

is optimal with worst case prediction error given by:

$$WPE\left[\phi^o\left(FSS^T\right)\right] = \frac{1}{2}\left[\overline{f}^*\left(\widetilde{w}^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right]$$

In order to show that $\phi_1^c\left(FSS^T\right)$ is almost optimal, first we note that from Theorem 2 in [1] we have that the functions $\overline{f}\,(w)$ and $\underline{f}\,(w)$ defined in (2) are the tightest lower and upper bound of $f_o(w)$, i.e.:

$$\begin{array}{l} \overline{f}\,(w) = \sup_{f \in FSS^T} f\,(w) \\ \underline{f}\,(w) = \inf_{f \in FSS^T} f\,(w) \end{array}$$

Thus, from (5) it follows that:

$$\underline{f}^*\left(\widetilde{w}^T\right) \leq \underline{f}\left(\widetilde{w}^T\right) \leq \phi_1^c\left(FSS^T\right) \leq \overline{f}\left(\widetilde{w}^T\right) \leq \overline{f}^*\left(\widetilde{w}^T\right)$$

From these inequalities and (6) it results:

$$\begin{array}{l} WPE\left[\phi_1^c\left(FSS^T\right)\right] = \\ = \max\left[\left|\phi_1^c\left(FSS^T\right) - \overline{f}^*\left(\widetilde{w}^T\right)\right|, \left|\phi_1^c\left(FSS^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right|\right] \leq \\ \leq \overline{f}^*\left(\widetilde{w}^T\right) - \underline{f}^*\left(\widetilde{w}^T\right) \equiv 2WPE\left[\phi^o\left(FSS^T\right)\right],\ \forall FSS^T \end{array}$$

i.e. the algorithm $\phi_1^c\left(FSS^T\right)$ is almost optimal.

In order to prove the remaining part of claim i), we now show that:

$$\begin{array}{l} \underline{f}_l\left(\widetilde{w}^T\right) \leq \underline{f}^*\left(\widetilde{w}^T\right) \\ \overline{f}_u\left(\widetilde{w}^T\right) \geq \overline{f}^*\left(\widetilde{w}^T\right) \end{array} \tag{7}$$

First, note that $\overline{w}^T$ and $\underline{w}^T$, defined in (4), are the solutions of the following optimization problems:

$$\begin{array}{l} \overline{w}^T = \arg\sup_{w \in B_\delta\left(\widetilde{w}^t\right)} \left\|w - \widetilde{w}^{\overline{t}}\right\| \\ \underline{w}^T = \arg\sup_{w \in B_\delta\left(\widetilde{w}^t\right)} \left\|w - \widetilde{w}_{\underline{t}}\right\| \end{array} \tag{8}$$

Suppose that $\overline{w}^T \in \overline{C}^{\overline{t}}$, then:

$$\begin{array}{rcl} \overline{f}^*\left(\widetilde{w}^T\right) & \equiv & \sup\limits_{w \in B_\delta\left(\widetilde{w}^T\right)} \overline{f}\,(w) = \overline{f}\left(\overline{w}^T\right) = \\ & = & \overline{f}\left(\widetilde{w}^T\right) + \gamma\delta^T \equiv \overline{f}_u\left(\widetilde{w}^T\right) \end{array}$$

At the contrary, if $\overline{w}^T \notin \overline{C}^{\overline{t}}$, we have:

$$\begin{array}{rcl} \overline{f}^*\left(\widetilde{w}^T\right) & \equiv & \sup\limits_{w \in B_\delta\left(\widetilde{w}^T\right)} \overline{f}\,(w) \leq \\ & \leq & \sup\limits_{w \in B_\delta\left(\widetilde{w}^T\right)} \left(\overline{h}^{\overline{t}} + \gamma\left\|w - \widetilde{w}^{\overline{t}}\right\|\right) = \\ & = & \overline{h}^{\overline{t}} + \gamma\left\|\overline{w}^T - \widetilde{w}^{\overline{t}}\right\| = \\ & = & \overline{h}^{\overline{t}} + \gamma\left\|\widetilde{w}^T - \widetilde{w}^{\overline{t}}\right\| + \gamma\delta^T \equiv \overline{f}_u\left(\widetilde{w}^T\right) \end{array}$$

Analogously it can be proven that $\underline{f}_l\left(\widetilde{w}^T\right) \leq \underline{f}^*\left(\widetilde{w}^T\right)$.

From (7), it follows that the worst-case prediction error of $\phi_1^c\left(FSS^T\right)$ is bounded as:

$$\begin{array}{l} WPE\left[\phi_1^c\left(FSS^T\right)\right] = \\ = \max\left[\left|\phi_1^c\left(FSS^T\right) - \overline{f}^*\left(\widetilde{w}^T\right)\right|, \left|\phi_1^c\left(FSS^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right|\right] \leq \\ \leq \max\left[\left|\phi_1^c\left(FSS^T\right) - \overline{f}_u\left(\widetilde{w}^T\right)\right|, \left|\phi_1^c\left(FSS^T\right) - \underline{f}_l\left(\widetilde{w}^T\right)\right|\right] = \\ = \frac{1}{2}\left[\overline{f}\left(\widetilde{w}^T\right) - \underline{f}\left(\widetilde{w}^T\right)\right] + \gamma\delta^T \end{array}$$

If $B_\delta\left(\widetilde{w}^T\right) \subseteq \overline{C}^t \cap \underline{C}^t$, then:

$$\begin{aligned}\overline{f}^*\left(\widetilde{w}^T\right) &\equiv \sup_{w\in B_\delta\left(\widetilde{w}^T\right)} \overline{f}\left(w\right) = \overline{f}\left(\widetilde{w}^T\right) + \gamma\delta^T \\ \underline{f}^*\left(\widetilde{w}^T\right) &\equiv \inf_{w\in B_\delta\left(\widetilde{w}^T\right)} \underline{f}\left(w\right) = \underline{f}\left(\widetilde{w}^T\right) - \gamma\delta^T\end{aligned}$$

$$\begin{aligned}WPE\left[\phi_1^c\left(FSS^T\right)\right] &= \frac{1}{2}\left[\overline{f}\left(\widetilde{w}^T\right) - \underline{f}\left(\widetilde{w}^T\right)\right] + \gamma\delta^T = \\ &= \frac{1}{2}\left[\overline{f}^*\left(\widetilde{w}^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right]\end{aligned}$$

Then, from (6), $y_c^{T+1} = \phi_1^c\left(FSS^T\right)$ is an optimal prediction with prediction error given by $\frac{1}{2}\left[\overline{f}\left(\widetilde{w}^T\right) - \underline{f}\left(\widetilde{w}^T\right)\right] + \gamma\delta^T$.

■

By exploiting more carefully the properties of the HVD $\overline{V}$ and $\underline{V}$, a somewhat stronger result can be obtained. Let $\overline{\mathcal{V}}(\widetilde{w}^T)$ the set composed of $\widetilde{w}^T$ and the vertices of the HVD $\overline{V}$ contained in $B_\delta\left(\widetilde{w}^T\right)$ and $\underline{\mathcal{V}}(\widetilde{w}^T)$ the set composed of $\widetilde{w}^T$ and the vertices of the HVD $\underline{V}$ contained in $B_\delta\left(\widetilde{w}^T\right)$. Let us define the following functions:

$$\overline{f}_l\left(\widetilde{w}^T\right) \doteq \begin{cases} \overline{f}_u\left(\widetilde{w}^T\right) & if\ \ \overline{w}^T \in \overline{C}^t \\ \max_{w\in\overline{\mathcal{V}}(\widetilde{w}^T)} \overline{f}\left(w\right) & otherwise \end{cases}$$

$$\underline{f}_u\left(\widetilde{w}^T\right) \doteq \begin{cases} \underline{f}_l\left(\widetilde{w}^T\right) & if\ \ \underline{w}^T \in \underline{C}^t \\ \min_{w\in\underline{\mathcal{V}}(\widetilde{w}^T)} \underline{f}\left(w\right) & otherwise \end{cases}$$

Theorem 3

i) The prediction algorithm:

$$\phi_2^c\left(FSS^T\right) = \frac{1}{2}\left[\underline{f}_u\left(\widetilde{w}^T\right) + \overline{f}_l\left(\widetilde{w}^T\right)\right]$$

is almost optimal, with prediction error bounded as:

$$\begin{gathered}\tfrac{1}{2}\left[\overline{f}_l\left(\widetilde{w}^T\right) - \underline{f}_u\left(\widetilde{w}^T\right)\right] \doteq \\ \doteq \underline{WPE}\left[\phi_2^c\left(FSS^T\right)\right] \leq \\ \\ \leq WPE\left[\phi_2^c\left(FSS^T\right)\right] \leq \\ \\ \leq \overline{WPE}\left[\phi_2^c\left(FSS^T\right)\right] \doteq \\ \doteq \max\left[\left|\phi_2^c\left(FSS^T\right) - \underline{f}_l\left(\widetilde{w}^T\right)\right|, \left|\phi_2^c\left(FSS^T\right) - \overline{f}_u\left(\widetilde{w}^T\right)\right|\right]\end{gathered}$$

ii) If $\underline{WPE}\left[\phi_2^c\left(FSS^T\right)\right] = \overline{WPE}\left[\phi_2^c\left(FSS^T\right)\right]$*, then the prediction* $y_c^{T+1} = \phi_2^c\left(FSS^T\right)$ *is optimal with minimal worst case prediction error given by:*

$$WPE\left[\phi_2^c\left(FSS^T\right)\right] = \frac{1}{2}\left[\overline{f}\left(\widetilde{w}^T\right) - \underline{f}\left(\widetilde{w}^T\right)\right] + \gamma\delta^T$$

Proof.

At first we show that:

$$\begin{gathered}\underline{f}_u\left(\widetilde{w}^T\right) \geq \underline{f}^*\left(\widetilde{w}^T\right) \\ \overline{f}_l\left(\widetilde{w}^T\right) \leq \overline{f}^*\left(\widetilde{w}^T\right)\end{gathered} \tag{9}$$

where $\underline{f}^*$ and $\overline{f}^*$ are given by (5) in the proof of theorem 2.

Suppose that $\overline{w}^T \in \overline{C}^{\overline{t}}$. From theorem 4 of [1] and from (8) in the proof of Theorem 2, it follows $\overline{f}\left(\overline{w}^T\right) = \sup_{w\in B_\delta\left(\widetilde{w}^{\overline{t}}\right)} \left(\overline{h}^{\overline{t}} + \gamma \left\|w - \widetilde{w}^{\overline{t}}\right\|\right)$. Since (2) implies $\overline{h}^{\overline{t}} + \gamma \left\|w - \widetilde{w}^{\overline{t}}\right\| \geq \overline{f}\left(w\right), \forall w \in W$, then $\overline{f}\left(\overline{w}^T\right) = \sup_{w\in B_\delta\left(\widetilde{w}^T\right)} \overline{f}\left(w\right)$. Thus, if $\overline{w}^T \in \overline{C}^{\overline{t}}$, we have:

$$\begin{aligned}\overline{f}^*\left(\widetilde{w}^T\right) &\equiv \sup_{w\in B_\delta\left(\widetilde{w}^T\right)} \overline{f}\left(w\right) = \overline{f}\left(\overline{w}^T\right) = \\ &= \overline{f}\left(\widetilde{w}^T\right) + \gamma\delta^T \equiv \overline{f}_u\left(\widetilde{w}^T\right)\end{aligned}$$

At the contrary, if $\overline{w}^T \notin \overline{C}^{\overline{t}}$, we have:

$$\begin{aligned}\overline{f}^*\left(\widetilde{w}^T\right) &\equiv \sup_{w\in B_\delta\left(\widetilde{w}^T\right)} \overline{f}\left(w\right) \geq \\ &\geq \max_{w\in\overline{\mathcal{V}}\left(\widetilde{w}^T\right)} \overline{f}\left(w\right) \equiv \overline{f}_l\left(\widetilde{w}^T\right)\end{aligned}$$

Analogously, it can be proven that $\underline{f}_u\left(\widetilde{w}^T\right) \geq \underline{f}^*\left(\widetilde{w}^T\right)$.

From (9) and from the definition of $\phi_2^c\left(FSS^T\right)$ it results that:

$$\underline{f}^*\left(\widetilde{w}^T\right) \leq \phi_2^c\left(FSS^T\right) \leq \overline{f}^*\left(\widetilde{w}^T\right)$$

Then, the worst-case prediction error of $\phi_2^c\left(FSS^T\right)$ is bounded as:

$$\begin{aligned}&WPE\left[\phi_2^c\left(FSS^T\right)\right] = \\ &= \max\left[\left|\phi_2^c\left(FSS^T\right) - \overline{f}^*\left(\widetilde{w}^T\right)\right|, \left|\phi_2^c\left(FSS^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right|\right] \leq \\ &\leq \overline{f}^*\left(\widetilde{w}^T\right) - \underline{f}^*\left(\widetilde{w}^T\right) \equiv 2WPE\left[\phi^o\left(FSS^T\right)\right],\ \forall FSS^T\end{aligned}$$

where the last equality follows from (6) in the proof of Theorem 2.

Thus the algorithm $\phi_2^c\left(FSS^T\right)$ is almost optimal.

From (9) it also follows that:

$$\begin{aligned}&WPE\left[\phi_2^c\left(FSS^T\right)\right] = \\ &\max\left[\left|\phi_2^c\left(FSS^T\right) - \overline{f}^*\left(\widetilde{w}^T\right)\right|, \left|\phi_2^c\left(FSS^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right|\right] \geq \\ &\geq \max\left[\left|\phi_2^c\left(FSS^T\right) - \overline{f}_l\left(\widetilde{w}^T\right)\right|, \left|\phi_2^c\left(FSS^T\right) - \underline{f}_u\left(\widetilde{w}^T\right)\right|\right] \geq \\ &\geq \tfrac{1}{2}\left[\overline{f}_l\left(\widetilde{w}^T\right) - \underline{f}_u\left(\widetilde{w}^T\right)\right] \equiv \underline{WPE}\left[\phi_2^c\left(FSS^T\right)\right]\end{aligned}$$

The lower bound of the prediction error of $\phi_2^c\left(FSS^T\right)$ is then derived. Now the upper bound is proved.

From (7) in the proof of Theorem 2 it follows that:

$$\begin{aligned}&WPE\left[\phi_2^c\left(FSS^T\right)\right] = \\ &= \max\left[\left|\phi_2^c\left(FSS^T\right) - \overline{f}^*\left(\widetilde{w}^T\right)\right|, \left|\phi_2^c\left(FSS^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right|\right] \leq \\ &\leq \max\left[\left|\phi_2^c\left(FSS^T\right) - \overline{f}_u\left(\widetilde{w}^T\right)\right|, \left|\phi_2^c\left(FSS^T\right) - \underline{f}_l\left(\widetilde{w}^T\right)\right|\right] \equiv \\ &\equiv \overline{WPE}\left[\phi_2^c\left(FSS^T\right)\right]\end{aligned}$$

Now suppose that

$$\underline{WPE}\left[\phi_2^c\left(FSS^T\right)\right] = \overline{WPE}\left[\phi_2^c\left(FSS^T\right)\right]$$

This happens only if:

$$\begin{gathered}\overline{f}_l\left(\widetilde{w}^T\right) = \overline{f}^*\left(\widetilde{w}^T\right) = \overline{f}_u\left(\widetilde{w}^T\right) \\ \underline{f}_l\left(\widetilde{w}^T\right) = \underline{f}^*\left(\widetilde{w}^T\right) = \underline{f}_u\left(\widetilde{w}^T\right)\end{gathered}$$

These equalities imply that:

$$WPE\left[\phi_2^c\left(FSS^T\right)\right] = \frac{1}{2}\left[\overline{f}^*\left(\widetilde{w}^T\right) - \underline{f}^*\left(\widetilde{w}^T\right)\right]$$

Then, from (6) in the proof of Theorem 2, $y_c^{T+1} = \phi_2^c\left(FSS^T\right)$ is an optimal prediction with prediction error given by:

$$\begin{aligned} WPE\left[\phi_2^c\left(FSS^T\right)\right] &= \frac{1}{2}\left[\overline{f}_l\left(\widetilde{w}^T\right) - \underline{f}_u\left(\widetilde{w}^T\right)\right] = \\ &= \frac{1}{2}\left[\overline{f}\left(\widetilde{w}^T\right) - \underline{f}\left(\widetilde{w}^T\right)\right] + \gamma\delta^T \end{aligned}$$

■

Remark 1

It can be noted that the condition $\underline{WPE}\left[\phi_1^c\left(FSS^T\right)\right] = \overline{WPE}\left[\phi_2^c\left(FSS^T\right)\right]$ under which the prediction $y_c^{T+1} = \phi_2^c\left(FSS^T\right)$ is optimal can actually be met. In particular, it is certainly met if $B_\delta\left(\widetilde{w}^T\right) \subseteq \overline{C}^t \cap \underline{C}^t$.

Remark 2

It can be proved that the worst-case prediction error of ϕ_2^c is better or equal than the one of ϕ_2^c, i.e. $WPE\left[\phi_2^c\left(FSS^T\right)\right] \leq WPE\left[\phi_1^c\left(FSS^T\right)\right]$. Moreover, the condition ii) in Theorem 3 for the prediction $\phi_2^c\left(FSS^T\right)$ to be actually optimal can be easily checked. These interesting features are paid at the expenses of an increase of computational complexity with respect to algorithm ϕ_1^c, since algorithm ϕ_2^c requires the computation of the vertices of the Hyperbolic Voronoi Diagrams $\overline{V}$ and $\underline{V}$.

■

In summary, the proposed NSM prediction method can be performed through a procedure consisting of a preliminary off-line phase, in which the parameters $\nu, \gamma, \varepsilon, \delta$ are chosen, and of an on-line phase, performed at each time step T, in which the prediction $\widehat{y}^{T+1}$ is evaluated.

Off-line phase: model calibration

1) Let $[1, T_{off}]$ be the time interval on which the off-line computation is performed and let the set of data recorded in this interval be called identification data set. Partition the identification data set in two parts. The first part, composed by data from 1 to $T_e < T_{off}$, called estimation data set, is used in steps 2 and 3. The second part, composed by data from $T_e + 1$ to T_{off}, called calibration data set, is used in step 4 for the selection of $\gamma, \varepsilon, \delta$ values.

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

3) Compute by means of theorem 1 the surface $\gamma^*(\varepsilon, \delta)$ defined by (3) on a suitable range of values of (ε, δ).

4) Select $(\gamma, \varepsilon, \delta)$ values in the validated region. A reasonable choice is $\widehat{\gamma} \cong \max_{v \in V}\|f_a'(w)\|$, $\widehat{\varepsilon} \cong$ accuracy of device used for y^t measurements and $\widehat{\delta}$ in the validated region, giving the minimum of $RMSE(\delta, \widehat{\gamma}, \widehat{\varepsilon})$, where $RMSE(\delta, \gamma, \varepsilon)$ is the prediction error of the predictor $\widehat{y}^{t+1} = \phi^c\left(FSS^{T_e}\right)$, $t \in [T_e + 1, T_{off}]$ obtained from theorem 2 or 3.

On-line phase: prediction

Suppose that data have been recorded up to time $T \geq T_{off}$. The NSM predictor of y^{T+1} is:

$$\widehat{y}^{T+1} = \phi^c \left(FSS^T \right)$$

where $\phi^c \left(FSS^T \right)$ is obtained from theorem 2 or 3 using the selected values $\widehat{\gamma}, \widehat{\varepsilon}, \widehat{\delta}$.

■

Some issues, regarding extensions and improvements of the proposed prediction method are now briefly discussed.

- In the on-line prediction, the time needed to evaluate $\widehat{y}^{T+1} = \phi^c \left(FSS^T \right)$ increases (linearly) with time T. As a consequence, for large values of T the predictions may be not computable within the required time. When this situation occurs a simple solution is to use a windowing technique making use of a constant number of data, excluding the oldest measurements. Note anyway that quite large amount of data can be actually processed before windowing is required. For example, the online computing time required for prediction using 10000 data is about 3 ms on a 2.8 GHz computer. Thus, the online computational burden may be an issue only for problems where the required times for prediction are very small.

- In the paper and in the above described procedure, given values of regression orders n_y, n_1, ... , 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 [20, 18, 30]. In the examples presented in section 4, 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 $RMSE\left(\widehat{\delta}, \widehat{\gamma}, \widehat{\varepsilon}\right)$.

- Using the same approach described above for one-step ahead prediction, almost optimal algorithms for k-step ahead prediction can be obtained using a model structure of the form:

  $$y^{t+k} = f\left(w^t\right)$$

  As typically done in statistical prediction methods, a k-step ahead prediction can be obtained by iterating k times the one-step ahead prediction. However, no optimality properties can be guaranteed for such a prediction.

- As it happens in other prediction methods, also in the present approach, selecting suitable scalings of regressors in order to adapt to the properties of data may prove very useful. An "optimal" scaling is proposed in Lemma 1 of [1].

- So far a global bound $\|f_o'(w)\| \leq \gamma$ over all W is assumed. However, a local approach, able to deal with a variable bound $\|f_o'(w)\| \leq \gamma(w)$ may be expected to give (possibly significant) improvements in prediction accuracy. 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. [31, 32, 33, 34]). However, finding such partitions may be not an easy task. A very simple alternative approach leading to variable gradient bound assumptions on f_o (see [1]), is based on the evaluation of a function f_a approximating f_o(using any desired method, e.g. the SM one proposed in this paper assuming global gradient bounds) and on the application of the method described in this paper to the residue function, defined as:

$$f_\Delta\left(w\right) \doteq f_o\left(w\right) - f_a\left(w\right)$$

using the set of values:

$$\Delta y^{t+1} = \widetilde{y}^{t+1} - f_a\left(w^t\right),\ t = 1, 2, ..., T-1$$

4 Numerical results

In this section three examples are presented. In the first one, prediction of the well known Wolf Sunspot Numbers series is considered. The prediction performances obtained by the proposed method for this real world series, widely used in the literature as a benchmark test, are compared with those provided by neural network predictors and by other linear and nonlinear predictors taken from the literature. In the second example, a predictor for the vertical acceleration of a quarter-car system with controlled suspensions is estimated. The predictor is tested in prediction over a time horizon required for model predictive control design. In the third example, prediction of time series generated by a linear AR system of order 2 driven by gaussian noise is considered. This example is aimed to evaluate the degree of conservativeness of the prediction performances of the NSM method, compared with those provided by optimal statistical predictors, which in this simulated example can be actually derived, since the models used for the statistical predictors have exactly the same structure of the data generation mechanism.

Example 1: Wolf Sunspot Numbers

The Wolf Sunspot Numbers time series shown in figure 3 is considered. This set of data, limited to year 1892, has been chosen because widely used as a benchmark in the time series literature [11, 12, 35].

The time series, consisting of 123 data, has been divided into an identification set, formed by the first 100 data and a forecasting set, composed by the remaining 23 data. The identification set has been used to estimate the predictors of the various methods. The forecasting set has been used to evaluate the prediction performances.

Nonlinear Set Membership Global predictor NSMG1_1

The one-step ahead predictor NSMG1_1 has been obtained assuming a model structure of the form:

$$y^{t+1} = f\left(w^t\right)$$
$$w^t = \left[y^t\ y^{t-1}\ y^{t-2}\ u^t\right]$$

with $u^t = 0$, $\varepsilon = 0$. It can be noted that better results could be obtained by using a higher number of regressors. We have made this choice to make a fair comparison with the methods presented below, which use a regressor order up to 4.

A global bound $\|f_o'(w)\| \leq \gamma$ on the norm of f_o' and an absolute noise bound δ have been considered. Then, the procedure described at the end of section 3 has been applied. The validation curve is shown

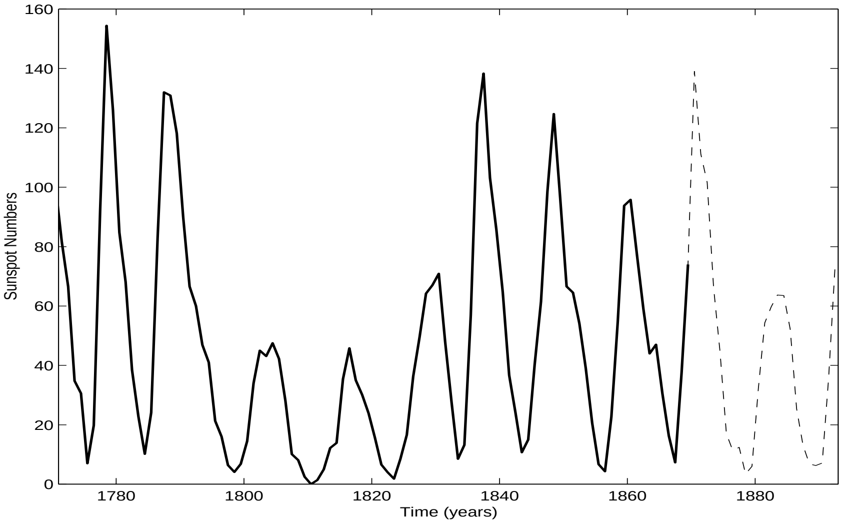

Figure 3: Example 1: Wolf sunspot numbers time series. 1770-1869 (bold line): estimation set. 1870-1892 (dshed line): forecasting set.

in figure 4. The values $\widehat{\delta} = 5$, $\widehat{\gamma} = 5.5$, have been chosen. It can be noted that the choice of these values is not very critical. This can be seen in figure 4 where the Root Mean Square Errors ($RMSE$) of the one-step ahead predictions on the forecasting set corresponding to different values of δ and γ are reported. The NSMG1_1 predictor is:

$$\widehat{y}^{T+1} = f_{G1}^{1}\left(\widetilde{w}^{T}\right) = \phi_1^c\left(FSS^T\right)$$

where $\phi_1^c\left(FSS^T\right)$ is computed from theorem 2, using the above values of γ and δ.

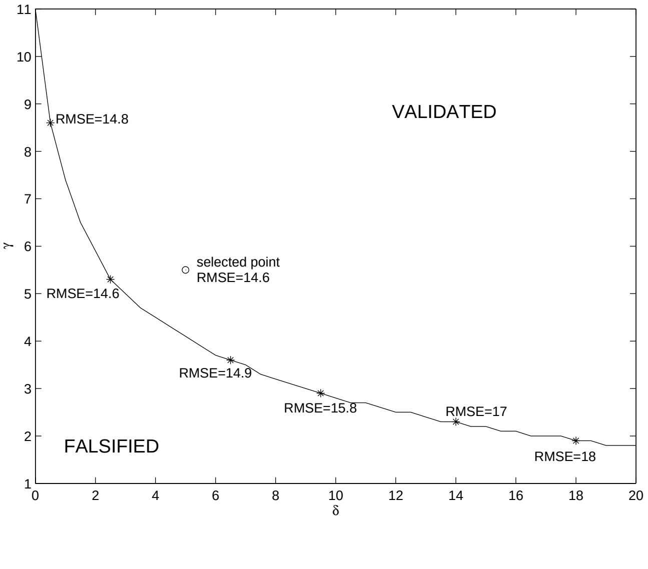

Figure 4: Example 1: validation curve for predictors NSMG1_1 and NSMG2_1.

Nonlinear Set Membership Global predictor NSMG2_1

The NSMG2_1 predictor is:

$$\widehat{y}^{T+1}=f_{G2}^{1}\left(\widetilde{w}^{T}\right)=\phi_{2}^{c}\left(FSS^{T}\right)$$

where $\phi_{2}^{c}\left(FSS^{T}\right)$ is computed from theorem 3, using the same assumptions as for predictor NSMG1_1.

Nonlinear Set Membership Local predictor NSML_1

The one-step ahead local predictor NSML_1 has been obtained assuming a model structure of the form:

$$\begin{aligned} y^{t+1} &= f\left(w^{t}\right) \\ w^{t} &= [y^{t}\ y^{t-1}\ u^{t}] \end{aligned}$$

with $u^{t}=0,\ \varepsilon=0$. The local approach described in [1] and at the end of section 3 has been used with $f_{a}\left(w^{t}\right)=f_{G1}^{1}\left(w^{t}\right)$. The procedure described at the end of section 3 has been applied to the residue time series $\widetilde{y}^{t+1}-f_{a}\left(\widetilde{w}^{t}\right),\, t=1,2,...,T-1$ assuming a global bound on the norm of $f_{\Delta}^{\prime}$ and an absolute noise bound δ. The regressors have been scaled as indicated in Lemma 1 of [1]. The values $\widehat{\delta}_{r}=1,\ \widehat{\gamma}_{r}=0.2$ have been chosen as indicated in step 4.

The NSML_1 predictor is:

$$\widehat{y}^{T+1}=f_{G1}^{1}\left(\widetilde{w}^{T}\right)+\phi_{1}^{c}\left(FSS^{T}\right)$$

where $\phi_{1}^{c}\left(FSS^{T}\right)$ is evaluated from theorem 2 by using the chosen values of δ_{r} and γ_{r}.

Nonlinear Set Membership Global predictor NSMG_11

The direct 11-step ahead Nonlinear Set Membership predictor NSMG_11 has been obtained assuming a model structure of the form:

$$\begin{aligned} y^{t+11} &= f\left(w^{t}\right) \\ w^{t} &= [y^{t}\ y^{t-2}\ y^{t-4}\ ...\ y^{t-12}\ u^{t}] \end{aligned}$$

with $u^{t}=0,\ \varepsilon=0$. A global bound $\|f_{o}^{\prime}(w)\|\leq\gamma$ on the norm of $f_{o}^{\prime}$ and an absolute noise bound δ have been considered. The procedure described at the end of section 3 has been applied, based on the evaluation of the validation curve $\gamma^{*}\left(\delta\right)$ shown in figure 5. The chosen values are $\widehat{\delta}=5,\ \widehat{\gamma}=7$. The NSMG_11 predictor is given by:

$$\widehat{y}^{T+11}=f_{G}^{11}\left(\widetilde{w}^{T}\right)=\phi_{1}^{c}\left(FSS^{T}\right)$$

where $\phi_{1}^{c}\left(FSS^{T}\right)$ is computed from theorem 2, using the above values of γ and δ.

Nonlinear Set Membership Local predictor NSML_11

The one-step ahead local predictor NSML_11 has been obtained assuming a model structure of the form:

$$\begin{aligned} y^{t+1} &= f\left(w^{t}\right) \\ w^{t} &= [y^{t}\ y^{t-2}\ y^{t-4}\ ...\ y^{t-12}\ u^{t}] \end{aligned}$$

with $u^{t}=0,\,\varepsilon=0$.

The local approach described in [1] and at the end of section 3 has been used with $f_{a}\left(w^{t}\right)=f_{G1}^{11}\left(w^{t}\right)$. The procedure described at the end of section 3 has been applied to the residue time series $\widetilde{y}^{t+11}-f_{a}\left(\widetilde{w}^{t}\right)$, $t=1,2,...,T-11$ assuming a global bound on the norm of $f_{\Delta}^{\prime}$ and an absolute noise bound δ. The

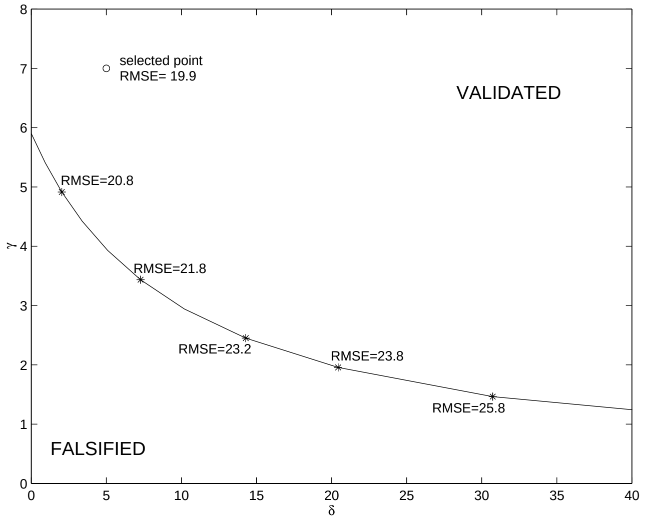

Figure 5: Example 1: validation curve for predictor NSMG_11.

regressors have been scaled as indicated in Lemma 1 of [1]. The chosen values are $\widehat{\delta}_r = 1$, $\widehat{\gamma}_r = 1$. The NSML_11 predictor is:

$$\widehat{y}^{T+11} = f_{G1}^{11}\left(\widetilde{w}^T\right) + \phi_1^c\left(FSS^T\right)$$

where $\phi_1^c\left(FSS^T\right)$ is evaluated from theorem 2, using the chosen values of δ_r and γ_r.

Neural Network predictor NN_1

The Neural Network predictor NN_1 has been obtained by considering a model of the form:

$$\begin{aligned} y^{t+1} &= \psi\left(w^t\right) \\ w^t &= [y^t\ y^{t-1}\ y^{t-2}] \end{aligned} \tag{10}$$

where the function ψ is a one hidden layer perceptron (see e.g. [36, 37]) composed by r neurons:

$$\psi\left(w\right) = \sum_{i=1}^{r} \alpha_i \sigma\left(\beta_i w - \lambda_i\right) + \zeta \tag{11}$$

Here $\alpha_i, \lambda_i, \zeta \in \Re$, $\beta_i \in \Re^n$, are parameters and $\sigma\left(x\right) = 2/(1+e^{-2x}) - 1$ is a sigmoidal function.

Several neural networks of the form (11) with different number of neurons (from $r = 3$ to $r = 20$) have been trained on the identification set. A neural network composed by $r = 8$ neurons, showing good performances in one-step ahead prediction, has been chosen for the predictor (10).

Neural Networks predictor NN_11

A Neural Network predictor NN_11, tuned for 11-step ahead prediction has been obtained by considering a model of the form:

$$\begin{aligned} y^{t+1} &= \psi\left(w^t\right) \\ w^t &= [y^t\ y^{t-1}\ ...\ y^{t-n_y+1}] \end{aligned} \tag{12}$$

where the function ψ is a neural network of the form (11). The 11-step ahead prediction is obtained by iterating the equation (12).

Several neural networks of the form (12) with different number of neurons (from $r = 3$ to $r = 20$) and different order of regression (from $n_y = 2$ to $n_y = 14$) have been trained on the identification set. A neural network with $r = 8$ neurons and with $n_y = 6$ regressors, showing good performances in multi-step ahead prediction, has been chosen for the predictor (12).

Several neural networks were also trained in order to obtain a direct 11-step ahead predictors of the form $y^{t+11} = \psi(w^t)$ but no satisfying results were obtained. This fact is probably due to the problem of local minima of neural networks risk function. Indeed such problem is usually relevant in case of strong nonlinearities and of large noise.

AR predictor [11]

This predictor is based on an autoregressive linear model of order 2.

Bilinear Predictor BL [35]

The BL predictor is based on a bilinear model.

SETAR predictor [38]

The SETAR predictor is based on a threshold autoregressive model.

Optimal Error Predictor OEP [21, 39]

The OEP predictor is obtained by a linear Set Membership approach.

Group Method of Data Handling predictor GMDH [40, 39]

GMDH is a nonlinear predictor, based on polynomial approximation. The original idea has been proposed in [40], and has been developed by several other authors. Here we show the results obtained in [39].

■

In table 2 the performance of the considered predictors on the forecasting set are reported. $RMSE_k$ and $MAXE_k$ with $k = 1, 11$ indicate the Root Mean Square Error and the Maximum Error in Absolute Value in the k-step ahead prediction.

Predictor	$RMSE_1$	$MAXE_1$	$RMSE_{11}$	$MAXE_{11}$
NSMG1_1	14.6	28	27.9	54
NSMG2_1	12.9	25	na	na
NSML_1	13.8	27	26.6	65
NSMG_11	na	na	19.9	45
NSML_11	na	na	17.7	45
NN_1	14.2	34	36.0	84
NN_11	na	na	23.4	63
AR	18.0	47	32.6	81
BL	16.6	46	32.6	81
SETAR	16.1	44	35.0	76
OEP	14.8	38	21.4	47
GMDH	14.7	42	29.4	81

Table 1. Example 1: one-step and 11-step ahead prediction errors

From these results, it can be noted that significant improvements are obtained by NSM predictors over the other methods, especially for 11-step ahead prediction.

It can be also noted that the performances of predictor NSMG2_1, making use of algorithm ϕ_2^c (theorem 3) are not significantly better than the ones of predictor NSMG1_1, making use of ϕ_1^c (theorem 2), suggesting that it may be not worth to use ϕ_2^c instead of the computationally simpler ϕ_1^c. Thus, in the next examples, only algorithm ϕ_1^c will be used.

Example 2: Quarter-car acceleration prediction

This example is related to prediction of vehicles vertical dynamics, an important tool in the automotive field, in view of the increasing diffusion of controlled suspension systems. Indeed, accurate prediction models may allow efficient control systems design through Model Predictive Control (MPC) methods.

Simulated data obtained by the quarter-car model with controlled semi-active suspensions shown in Figure 6 have been considered in this example. NSM identification of half-car model using experimental data is reported in [41] and [42].

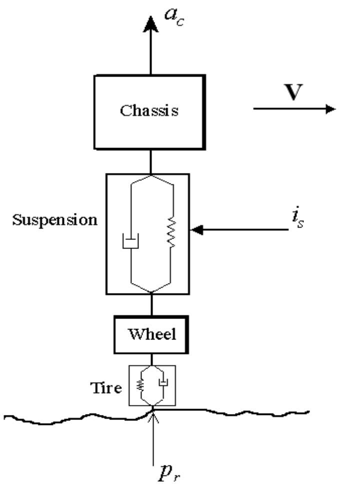

Figure 6: Example 2: The quarter-car model.

The quarter-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: road profile p_r, suspension control current i_s, chassis vertical acceleration a_c. It is considered that the road profile $p_r(t)$ is known, that current $i_s(t)$ can be measured with a precision of 3.75% and that the variable $a_c(t)$ can be measured with a precision of 5%.

The chassis, is simulated as rigid body. The following static nonlinear characteristic has been assumed for the tire:

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

where F_1 is the tire force, Δp_1 and Δv_1 are the differences of position and velocity at the extremes of tire, $\beta_1 = 10000\ Ns/m$ and $F_{1E}(\Delta p_1)$ is shown in Figure 7b. The following nonlinear characteristic has been assumed for the controlled suspension:

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

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 Figure 7a for the two extreme values $i = 0\ A$ and $i = 1.6\ A$.

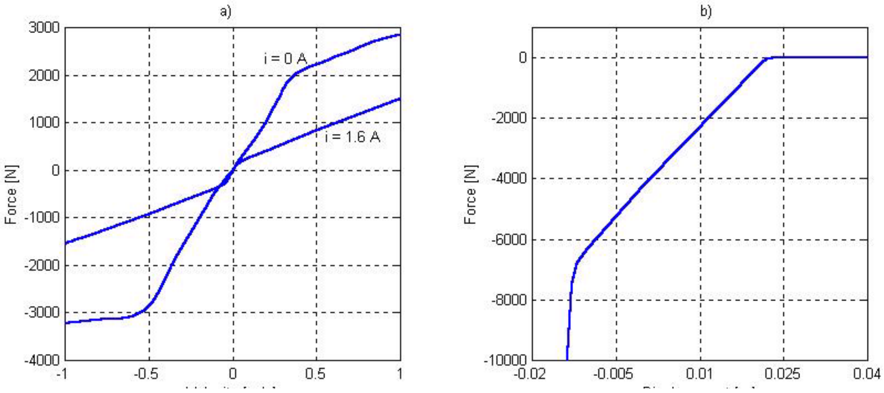

Figure 7: Example 2. a) Force-velocity chracteristic F_{2D} of suspension. b) Force-displacement characteristic F_{1E} of tires.

A data set has been generated from "true system" simulation, for a period of 24 seconds, using a random profile with amplitude $\leq$ 4 cm. The data set consists of the values of p_r, i_s and a_c recorded with a sampling time of $\tau = 1/512$ sec. The sequence of each measured variables is composed of 12280 samples. The values of a_c have been corrupted by uniformly distributed noises of relative amplitude 5% and the values of i_s have been corrupted by uniformly distributed noises of relative amplitude 5%. The data set related to the first 20 seconds, called identification data set, has been used for off-line procedure. The data set related to the last 4 seconds, called validation data set, has been used to test the prediction 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, [42].

Two predictor, relating front chassis accelerations to the road profile at the sampling times, have been obtained from the identification data set considering a model of the form:

$$y^{t+1} = f(w^t)$$
$$w^t = [y^t \ ... \ y^{t-3} \ u_1^t \ u_1^{t-1} \ u_2^t \ ]$$

with $y^t = a_c(t\tau)$, $u_1^t = p_r(t\tau)$ and $u_2^t = i_s(t\tau)$. The regressors orders $n_y = 4$, $n_1 = 2$, $n_2 = 1$ have been chosen as described at the end of section 3.

Nonlinear Set Membership Local predictor NSML

The predictor, called NSML, has been derived by means of the local approach mentioned at the end of section 3 with $f_a(w) = \theta^T w$, where θ is the coefficients vector of an ARX model estimated by means of the least squares method. The procedure described in section 3 has been applied to the residue data $\Delta y^{t+1} = \widetilde{y}^{t+1} - \theta^T \widetilde{w}^t$, $t = 1, 2, ..., 12280$. The 7680 data corresponding to the first 15 seconds of the identification set have been taken as estimation set, the 1536 data corresponding to the last 5 seconds have been taken as calibration set.

The regressors have been scaled, according to Lemma 1 in [1]. A bound $\|f'_\Delta(w)\| \leq \gamma_r$ on the gradient of residue function $f_\Delta(w) = f_o(w) - f_a(w)$ and a relative noise bound $|d^t| \leq \varepsilon_r |\Delta y^{t+1}| + \gamma_r \delta_r \|w^t\|$, $\forall t$, have been assumed. The values $\widehat{\varepsilon}_r = 0.2$, $\widehat{\delta}_r = 1.2$ and $\widehat{\gamma}_r = 0.67$ have been chosen, according to the procedure described in steps 3 and 4 of the off-line phase and using the residue data Δy^t. The NSML predictor is based on the model:

$$y^{t+1} = f_a\left(w^t\right) + \phi_1^c\left(FSS^t\right)$$

where $\phi_1^c\left(FSS^t\right)$ is evaluated from theorem 2 by using the chosen values of ε_r, δ_r and γ_r.

Neural Network predictor NN

The Neural Network predictor NN has been obtained by considering a one hidden layer perceptron (see e.g. [36, 37]) of the form (11). Several neural networks of such form with different number of neurons (from $r = 3$ to $r = 20$) have been trained on the identification set. A neural network composed by $r = 6$ neurons, showing good performances in multi-step ahead prediction, has been chosen.

$\blacksquare$

The NSML and NN predictors have been tested on the validation set, evaluating their predictive ability over the time horizon that may be required for MPC design, i.e. $k = [1, 150]$. In Figure 8, a typical comparison between "true" data and of the ones predicted by the NSML and NN predictors are reported, showing that NSML model can be reliably used for MPC design.

Example 3: Is the SM approach too conservative?

As argumented in the paper, and confirmed by the previous examples, the fact that the SM approach makes use of quite weak assumptions may be a key feature for cases where reliable information on the regression function and on noise is not available. However, it may be guessed that the SM approach could give very conservative results when reliable information on the parametric structure of the system and on noise statistics is available.

In order to have some hints on this point, a most adverse situation for SM approach is simulated in this example: data are generated by a linear AR model driven by i.i.d. gaussian white noise and the NSM predictors are compared with the optimal statistical predictor, which makes use of the exact assumptions.

A Montecarlo simulation composed by 200 time series of 150 data was generated by the equation:

$$y^{t+1} = 14.7 + 1.425 y^t - 0.731 y^{t-1} + d^t$$

where d^t is a gaussian white noise of variance $\sigma_d^2 = 50$. Note that this is the AR model derived in [11] for the prediction of the Wolf Sunspot Numbers. The time series have been divided into an identification set of 100 data, used for constructing the different predictors, and a forecasting set of 50 data, used for their prediction accuracy evaluation. For each time series, the following predictors have been derived.

Nonlinear Set Membership Global predictor NSMG_1

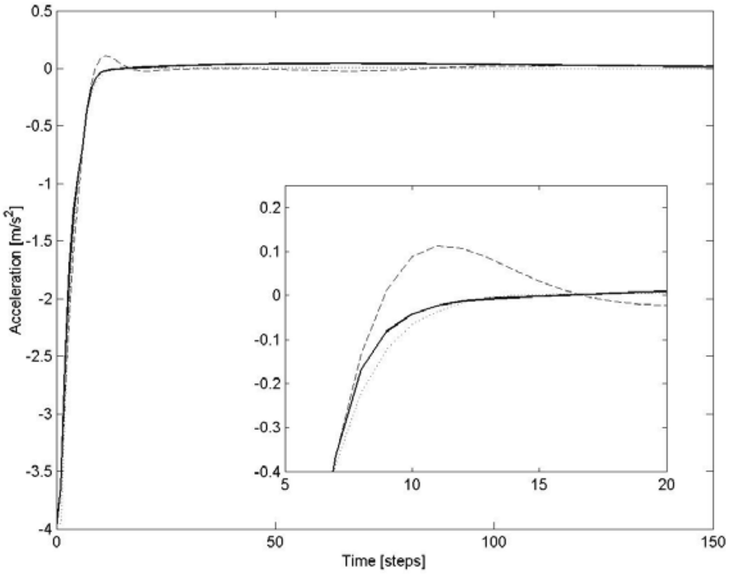

Figure 8: Example 2. Chassis accelerations: "true" (solid line), predicted by NSML model (dotted line), predicted by NN model (dashed line).

The one-step ahead predictor NSMG_1 has been obtained assuming a model structure of the form:

$$y^{t+1} = f\left(w^{t}\right)$$
$$w^{t} = [y^{t}\; y^{t-1}\; u^{t}]$$

with $u^{t} = 0,\ \varepsilon = 0$.

A global bound $\|f_o'(w)\| \leq \gamma$ on the norm of f_o' and an absolute noise bound δ have been considered. The procedure described at the end of section 3 has been applied and the values $\widehat{\delta} = 15$, $\widehat{\gamma} = 2$, have been chosen. The NSMG_1 predictor is:

$$\widehat{y}^{T+1} = f_G^1\left(\widetilde{w}^T\right) = \phi_1^c\left(FSS^T\right)$$

where $\phi_1^c\left(FSS^t\right)$ is computed from theorem 2, using the above values of γ and δ.

Nonlinear Set Membership Global predictor NSMG_11

The direct 11-step ahead predictor NSMG_11 has been obtained assuming a model structure of the form:

$$y^{t+11} = f\left(w^{t}\right)$$
$$w^{t} = [y^{t}\; y^{t-1}\; u^{t}]$$

with $u^{t} = 0,\ \varepsilon = 0$.

A global bound $\|f_o'(w)\| \leq \gamma$ on the norm of f_o' and an absolute noise bound δ have been considered. The procedure described at the end of section 3 has been applied. The chosen values are: $\widehat{\delta} = 13$, $\widehat{\gamma} = 2.1$. The NSMG_11 predictor is:

$$\widehat{y}^{T+11} = f_G^{11}\left(\widetilde{w}^T\right) = \phi_1^c\left(FSS^T\right)$$

where $\phi_1^c\left(FSS^T\right)$ is computed from theorem 2, using the above values of δ and γ.

Nonlinear Set Membership Local predictor NSML_1

The one-step ahead local predictor NSML_1 has been obtained assuming a model structure of the form:

$$\begin{aligned} y^{t+1} &= f\left(w^t\right) \\ w^t &= \left[y^t \; y^{t-1} \; u^t\right] \end{aligned}$$

with $u^t = 0, \; \varepsilon = 0$.

The NSML_1 predictor has been derived using the local approach mentioned at the end of section 3 with $f_a\left(w^t\right) = f_G^1\left(w^t\right)$. The procedure of section 3 has been applied to the residue time series $\widetilde{y}^{t+1} - f_a\left(\widetilde{w}^t\right), \; t = 1, 2, ..., T-1$ assuming a global bound on the weighted norm of f'_Δ and a absolute noise bound δ. The regressors have been scaled as indicated in Lemma 1 of [1]. The values $\widehat{\delta}_r = 5$, $\widehat{\gamma}_r = 0.3$ have been chosen. The NSML_1 predictor is:

$$\widehat{y}^{T+1} = f_G^1\left(\widetilde{w}^T\right) + \phi_1^c\left(FSS^T\right)$$

where $\phi_1^c\left(FSS^T\right)$ is evaluated from theorem 2 by using the chosen values of δ_r and γ_r.

Nonlinear Set Membership Local predictor NSML_11

The direct 11-step ahead predictor NSML_11 was obtained assuming a model structure of the form:

$$\begin{aligned} y^{t+11} &= f\left(w^t\right) \\ w^t &= \left[y^t \; y^{t-1} \; u^t\right]^T \end{aligned}$$

with $u^t = 0, \; \varepsilon = 0$.

This predictor has been derived using the local approach mentioned at the end of section 3 with $f_a\left(w^t\right) = f_G^{11}\left(w^t\right)$. The procedure described in section 3 has been applied to the residue time series $\widetilde{y}^{t+11} - f_a\left(\widetilde{w}^t\right), \; t = 1, 2, ..., T-1$ assuming a global bound on the weighted norm of f'_Δ. The regressors have been scaled as indicated in Lemma 1 of [1]. The values $\widehat{\delta}_r = 4$, $\widehat{\gamma}_r = 0.3$ have been chosen. The NSML_11 predictor is:

$$\widehat{y}^{T+11} = f_G^{11}\left(\widetilde{w}^T\right) + \phi_1^c\left(FSS^T\right)$$

where $\phi_1^c\left(FSS^T\right)$ is evaluated from theorem 2 by using the chosen values of δ_r and γ_r.

Linear Autoregressive predictor AR

This predictor is based on a linear AR model the same structure of the system generating the data:

$$y^{t+1} = a_0 + a_1 y^t - a_2 y^{t-1} + d^t$$

The parameters a_0, a_1 and a_2 and the variance of d^t have been derived on the identification set using the Matlab Identification Toolbox.

■

In table 1 the results obtained from the 200 experiments are reported. The root mean square errors and the maximum prediction errors in absolute value evaluated on the forecasting set and averaged over the 200 realizations of the time series are indicated with $\overline{RMSE}_k$ and $\overline{MAXE}_k$ for the k-step ahead prediction, with $k = 1, 11$.

Predictor	$\overline{RMSE}_1$	$\overline{MAXE}_1$	$\overline{RMSE}_{11}$	$\overline{MAXE}_{11}$
NSMG_1	8.9	23	19.2	46
NSMG_11	-	-	19.4	45
NSML_1	8.5	22	19.5	47
NSML_11	-	-	18.8	43
AR	7.2	18	17.9	43

Table 1. Example 3: one-step and 11-step prediction error averages.

In this simulation example, the AR predictor gives "optimal" one-step ahead prediction performance, since the data are actually generated by the linear model structure and statistical noise assumptions used for model identification. The results of table 1 shows the interesting fact that the NSM predictors, though not using such strong informations on data generating mechanism, do not exhibit a significant deterioration in the prediction performances.

5 Conclusions

In the paper, a prediction method for nonlinear time series has been presented, based on a Set Membership approach, requiring quite weak assumptions on noise and on involved nonlinearities. At difference with most of methods in the literature, the NSM method does not require to know the functional form of nonlinearities, thus reducing the effects of modeling errors on the propagation of prediction errors, allowing to increase the horizon on which reliable predictions can be made. 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 in many applications and anyway is lost in presence of approximate modeling. On the base of these theoretical features, it is expected that the proposed predictors may have good performances, especially for the multi-step ahead prediction, and exhibit good robustness versus imprecise knowledge of involved nonlinearities and of noise properties. These expectations appear to be confirmed by the presented numerical results. Indeed, in the case of the Wolf Sunspot Numbers series, widely used in the literature as a benchmark test, the NSM predictors display sensible improvements over linear and nonlinear predictors taken from the literature, in particular for multistep ahead prediction. On the other hand, the simulated linear example with i.i.d. gaussian noise shows surprisingly small performance deteriorations of NSM predictors with respect to optimal statistical ones, which in this case can be derived, since the used models have the same structure of the data generation mechanism.

Further applications of the NSM prediction method to real data can be found in [43], [41], [44] where prediction of river flow, half-car acceleration and troposphere pollutants are considered, respectively.

In conclusion, the new approach to nonlinear time series prediction presented in this paper appears to give quite promising results and is being tested on larger classes of simulated and real problems.

References

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

[2] B. Widrow and S. D. Stearns, *Adaptive Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1985.

[3] L. Wu, M. Niranjan, and F. Fallside, "Fully vector-quantized neural network-based code-excited nonlinear predictive speech coding," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 482–489, 1994.

[4] B. Y. Ryabko, "Prediction of random sequences and universal coding," *Probl. Inform. Transm.*, vol. 24, pp. 87–96, 1988.

[5] J. M. Hutchinson, A. W. Lo, and T. Poggio, "A nonparametric approach to pricing and hedging derivative securities via learning networks," *J. Finance*, vol. XLIX, no. 3, pp. 677–687, 1994.

[6] A. S. Weigend and N. A. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, 1994.

[7] A. Porporato and L. Ridolfi, "Nonlinear analysis of river flow time sequences," *Water Resour. Res., 33(6)*, pp. 1353–1367, 1997.

[8] A. W. Jayawardena and A. B. Gurung, "Noise reduction and prediction of hidrometereological time series," *J. Hydrol.*, vol. 227, pp. 242–264, 2000.

[9] B. Sivakumar, "Chaos theory in hydrology: important issues and interpretations," *J. Hydrol.*, vol. 227(1-4), pp. 1–20, 2000.

[10] J. Houghton, "The bakerian lecture, 1991: The predictability of weather and climate," *Phil. Trans. Roy. Soc. London A*, vol. 337, pp. 521–572, 1991.

[11] G. Box and G. Jenkins, *Time series analysis: Forecasting and control.* San Francisco, CA: Holden Day, 1976.

[12] G. M. Jenkins, *Practical Experiences with Modelling and Forecasting Time Series.* Time Series Library, G.J.P. England, 1979.

[13] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica, 36(6)*, pp. 789–814, 2000.

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

[15] D. S. Modha and E. Masry, "Memory-universal prediction of stationary random processes," *IEEE Transaction on Information Theory*, vol. 44, pp. 117–133, 1998.

[16] K. R. Muller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Advances in Kernel Methods — Support Vector Learning*, (Cambridge, MA), pp. 243–254, 1999.

[17] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems.* Winston, Washington DC, 1977.

[18] 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.

[19] A. Juditsky, H. Hjalmarsson, A. Benveniste, B.Delyon, L. Ljung, J. Sjöberg, and Q. Zhang, "Nonlinear black-box modeling in system identification: Mathematical foundations," *Automatica*, vol. 31, pp. 1725–1750, 1995.

[20] R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic systems - a survey on input/output approaches," *Automatica*, vol. 26, pp. 651–677, 1990.

[21] M. Milanese and R. Tempo, "Optimal algorithms theory for robust estimation and prediction," *IEEE Transaction on Automatic Control*, vol. 30, pp. 730–738, 1985.

[22] 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.

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

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

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

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

[27] A. Pinkus, *n-Widths in Approximation Theory.* Berlin: Springer-Verlag, 1985.

[28] G. W. Wasilkowski and H. Woźniakowski, "Complexity of weighted approximation over R^d," *Journal of Complexity*, no. 17, pp. 722–740, 2001.

[29] H. Edelsbrunner, *Algorithms in Combinatorial Geometry.* Berlin: Springer-Verlag, 1987.

[30] K. S. Narendra and S. Mukhopadhyay, "Neural networks for system identification," in *Sysid 97*, vol. 2, pp. 763–770, 1997.

[31] A. Stenman, F. Gustafsson, and Ljung, "Just in time models for dynamical systems," in *Proc. of the 35th IEEE Conference on Decision and Control*, (Kobe, Japan), pp. 1115–1120, 1996.

[32] Q. Zheng and H. Kimura, "Just in time modeling for function prediction and its applications," *Asian Journal of Control, Vol. 3, No. 1,*, pp. 35–44, 2001.

[33] E. D. Sontag, "Nonlinear regulation. The piecewise linear approach," *IEEE Trans. Automatic Control*, vol. 26, pp. 346–357, 1981.

[34] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari, "A clustering technique for the identification of piecewise affine systems," *A. Sangiovanni-Vincentelli and M.D. Di Benedetto, Eds., Hybrid Systems: Computation and Control, Lecture Notes in Computer Science. Springer Verlag*, 2001.

[35] C. W. J. Granger and A. P. Andersen, "Introduction to bilinear time series models," *Vandenhoek and Ruprecht, Gottingen*, 1978.

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

[37] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer Verlag, 1995.

[38] H. Tong and K. S. Lim, "Threshold autoregression, limit cycles and cyclical data," *Journal of the Royal Statistical Society, Series B 42*, pp. 245–292, 1980.

[39] A. Vicino, R.Tempo, R. Genesio, and M. Milanese, "Optimal error and GMDH predictors," *International Journal of Forecasting*, no. 3, pp. 313–328, 1987.

[40] A. G. Ivakhnenko, "Heuristic self-organization in problems of engineering cybernetics," *Automatica*, vol. 6, pp. 207–219, 1970.

[41] M. Milanese, C. Novara, F. Mastronardi, and D. Amoroso, "Experimental modeling of vertical dynamics of vehicles with controlled suspensions," in *SAE World Congress*, (Detroit, Michigan), 2004.

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

[43] M. Milanese and C. Novara, "Set Membership prediction of river flow," *Systems and Control Letters*, vol. 53/1, pp. 31–39, 2004.

[44] G. Finzi, M. Milanese, C. Novara, and M. Volta, "Nonlinear Set Membership forecast of urban ozone peaks," in *IFAC 2005*, (Prague, Czech Republic), 2005.