TABLE III
SIMULATION ERRORS FOR THE SILVERBOX DATA, OVER THE FULL TEST SET (CASE I) AND ONLY FOR THE LAST 10 000 POINTS OF THE TEST SET (CASE II)

Case	Indicators	Black-Box model	Partially Linear Model
Case I	$\max(\lvert e_i\rvert)$	0.0081	0.0037
	$\text{mean}(\lvert e_i\rvert)$	2.30×10^{-4}	2.02×10^{-4}
	$\text{RMSE}(e_i)$	3.24×10^{-4}	2.71×10^{-4}
Case II	$\max(\lvert e_i\rvert)$	0.0081	0.0037
	$\text{mean}(\lvert e_i\rvert)$	3.72×10^{-4}	2.31×10^{-4}
	$\text{RMSE}(e_i)$	5.86×10^{-4}	3.34×10^{-4}

mean squared error decreases to 2.7×10^{-4}. Moreover, when considering only the last 10 000 points of the test data, the improvement is more important, as shown in Table III. Using the full black-box model, the maximum absolute error is 0.0081, which is reduced to 0.0037 with the PL-LSSVM. The mean absolute error for the full black-box model is 2.3×10^{-4}; for the partially linear model, it is 2.02×10^{-4}. The effective number of parameters is reduced from 490 to 190.

V. CONCLUSION

In this note, we illustrated that it is possible to use a partially linear model with least squares support vector machines to successfully identify a model containing a linear part and a nonlinear component, with better performance results than a full nonlinear black-box model. The structured model may show a better generalization ability, and a reduced effective number of parameters, than a full nonlinear black-box model. In the real-life example of the Silverbox benchmark data, an existing nonlinear black-box model can be further improved by imposing a linear structure, as it is illustrated in the simulation performance.

REFERENCES

[1] M. Espinoza, K. Pelckmans, L. Hoegaerts, J. A. K. Suykens, and B. De Moor, "A comparative study of LS-SVM's applied to the silverbox identification problem," in *Proc. 6th IFAC Symp. Nonlinear Control Systems (NOLCOS)*, 2004, pp. 513–518.

[2] M. Espinoza, J. A. K. Suykens, and B. De Moor, "Partially linear models and least squares support vector machines," in *Proc. 43rd IEEE Conf. Decision and Control*, 2004, pp. 3388–3393.

[3] J. Gao, "Semiparametric nonlinear time series model selection," *J. R. Statist. Soc. B*, vol. 66, pp. 321–336, 2004.

[4] S. R. Gunn and J. S. Kandola, "Structural modeling with sparse kernels," *Mach. Learn.*, vol. 48, pp. 137–163, 2002.

[5] W. Härdle, H. Liang, and J. Gao, *Partially Linear Models*. Heidelberg, Germany: Physica-Verlag, 2000.

[6] S. Haykin, *Neural Networks, a Comprehensive Foundation*. New York: Macmillan, 1994.

[7] D. Lindgren, "Projection techniques for classification and identification," Ph.D. dissertation, Linköping Univ., Linköping, Sweden, 2005.

[8] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 1987.

[9] C. L. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, pp. 661–675, 1973.

[10] P. M. Robinson, "Root n-consistent semiparametric regression," *Econometrica*, vol. 56, no. 4, pp. 931–954, 1988.

[11] J. Schoukens, G. Nemeth, P. Crama, Y. Rolain, and R. Pintelon, "Fast approximate identification of nonlinear systems," *Automatica*, vol. 39, no. 7, 2003.

[12] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson, and A. Juditsky, "Nonlinear black-box modeling in system identification: A unified overview," *Automatica*, vol. 31, pp. 1691–1724, 1995.

[13] P. Speckman, "Kernel smoothing in partial linear models," *J. R. Statist. Soc. B*, vol. 50, 1988.

[14] J. A. K. Suykens, T. Van Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[15] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[16] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.

[17] C. K. I. Williams and M. Seeger, "Using the nystrom method to speed up kernel machines," in *Proc. NIPS 2000*, vol. 13, V. Tresp, T. Leen, and T. Dietterich, Eds., Vienna, Austria, 2000.

[18] Y. Yu and W. Lawton, "Wavelet based modeling of nonlinear systems," in *Nonlinear Modeling: Advanced Black Box Techniques*, J. A. K. Suykens and J. Vandewalle, Eds. Norwell, MA: Kluwer, 1998, pp. 119–148.

Model Quality in Identification of Nonlinear Systems

Mario Milanese and Carlo Novara

***Abstract*—In this note, the problem of the quality of identified models of nonlinear systems, measured by the errors in simulating the system behavior for future inputs, is investigated. Models identified by classical methods minimizing the prediction error, do not necessary give "small" simulation error on future inputs and even boundedness of this error is not guaranteed. In order to investigate the simulation error boundedness (SEB) property of identified models, a Nonlinear Set Membership (NSM) method recently proposed by the authors is taken, assuming that the nonlinear regression function, representing the difference between the system to be identified and a linear approximation, has gradient norm bounded by a constant γ. Moreover, the noise sequence is assumed unknown but bounded by a constant ε. The NSM method allows to obtain validation conditions, useful to derive "validated regions" within which to suitably choose the bounding constants γ and ε. Moreover, the method allows to derive an " optimal" estimate of the true system. If the chosen linear approximation is asymptotically stable (a necessary condition for the SEB property), in the present note a sufficient condition on γ is derived, guaranteeing that the identified optimal NSM model has the SEB property. If values of γ in the validated region exist, satisfying the sufficient condition, the previous results can be used to give guidelines for choosing the bounding constants γ and ε, additional to the ones required for assumptions validation and useful for obtaining models with "low" simulation errors. The numerical example, representing a mass-spring-damper system with nonlinear damper and input saturation, demonstrates the effectiveness of the presented approach.**

***Index Terms*—Identification, nonlinear systems, Set Membership, simulation error, stability.**

I. INTRODUCTION

Consider a nonlinear dynamic system of the form

$$y_{t+1} = f_o(w_t) = f_o(x_t, v_t) \tag{1}$$

where: $w_t = (x_t, v_t), x_t = [y_t \dots y_{t-n+1}]^T, v_t = [u_t^1 \dots u_{t-n_1+1}^1 \dots u_t^q \dots u_{t-n_q+1}^q]^T, y_t, u_t^1, \dots, u_t^q \in \mathbb{R}, m = n + n_u, n_u = \sum_{i=1}^{q} n_i, f_o : \mathbb{R}^m \to \mathbb{R}$, and $t = 0, 1, 2, \dots$.

Suppose that the function f_o is not known, but a set of noise corrupted measurements $\widetilde{y}_t$ and $\widetilde{v}_t$, of y_t and $v_t, t = 0, 1, 2, \dots, T$ is available. Then, the aim is to find an estimate $\widehat{f}$ of f_o such that the simulation error for future input sequence is "small."

Most of identification methods in the literature (see, e.g., [1]–[4]) consider that f_o belongs to a finitely parametrized set of functions

Manuscript received May 28, 2004; revised March 15, 2005 and June 1, 2005. Recommended by Guest Editor L. Ljung.

The authors are with the Dipartimento di Automatica e Informatica, Politecnico di Torino, 10129 Torino, Italy (e-mail: mario.milanese@polito.it; carlo.novara@polito.it).

Digital Object Identifier 10.1109/TAC.2005.856657

$K \doteq \{f_p(\varphi), p \in \mathbb{R}^r, \varphi \in \mathbb{R}^m\}$ and measured data are used to derive an estimate $\widehat{p}$ of p, leading to an estimate $\widehat{f} = f_{\widehat{p}}$. The estimate $\widehat{p}$ of p is usually obtained by minimizing an error function $\widehat{p} = \arg\min_p V(p, \Phi_T), V(p, \Phi_T) = \sum_{t=0}^{T-1} |\widetilde{y}_{t+1} - f_p(\varphi_t)|^l$, where φ_t is a regression vector and $\Phi_T = [\varphi_0, \varphi_1, \ldots, \varphi_T]$. Several choices can be taken for the regressor φ_t. Widely used are the following ones:

$$\varphi_t = \widetilde{w}_t = [\widetilde{y}_t \ \cdots \ \widetilde{y}_{t-n+1} \ \widetilde{v}_t]^T$$
$$\varphi_t = \widehat{w}_t = [f_p(\widehat{w}_{t-1}) \ \cdots \ f_p(\widehat{w}_{t-n}) \ \widetilde{v}_t]^T$$

leading to NARX and NOE models, respectively (see, e.g., [2] and [5]).

Such an approach is often indicated as prediction error (PE) method, since $V(p, \Phi_T)$ is an estimate of the prediction error for the given regressor choice. A basic problem is that an estimated model $\widehat{f}$ giving low prediction error, does not necessary give low simulation errors on future inputs (see, e.g., [2]). Indeed, even boundedness of the simulation error of $\widehat{f}$ is not guaranteed. The problem arises even in the identification of linear systems. In fact, the PE estimate of a linear asymptotically stable system may result to be not asymptotically stable, in which case the simulation error is generally unbounded. If in the linear case it can be anyway *a posteriori* detected if the estimated model $\widehat{f}$ is not guaranteed to have bounded simulation errors, the problem is more difficult in the nonlinear case. In order to formulate more precisely the simulation error boundedness (SEB) problem investigated in this note, let $y_t(f, x_0, v), t = 0, 1, 2, \ldots$ be the solution of system $y_{t+1} = f(x_t, v_t)$, corresponding to initial condition $x_0 = [y_0 \ \ldots \ y_{-n+1}]^T$ and input sequence $v = [v_0, v_1, v_2, \ldots]$. For given initial condition x_0 and input sequence v, the simulation error at time t of the identified model $\widehat{f}$ is given by

$$SE_t(f_o, \widehat{f}) \doteq y_t(f_o, x_0, v) - y_t(\widehat{f}, \widetilde{x}_0, v)$$

where $\widetilde{x}_0 = [\widetilde{y}_0 \cdots \widetilde{y}_{-n+1}]^T$. Following a similar definition in [6], we say that model $\widehat{f}$ simulates system f_o on the sets K_1, K_2 in time $\overline{T}$ and up to accuracy δ if for each $x_0 \in K_1$ and $v \in K_2$ it results

$$|SE_t(f_o, \widehat{f})| \leq \delta, \qquad t = 0, 1, 2, \ldots, \overline{T}. \tag{2}$$

Some results are available showing that choosing a neural net as parametrized function $f_p(\varphi), p \in \mathbb{R}^r$ with sufficiently large r, a $\widehat{f}$ exists which simulates system f_o up to any given accuracy δ on compact sets K_1, K_2, in any finite time $\overline{T}$ for which the solutions of (1) exist [6], [7], and in infinite time in case of fading memory systems [8]. However, in practical applications it is important to evaluate if the actually identified model $\widehat{f}$ (obtained by the chosen identification method operating on the available noise corrupted data) simulates system f_o up to finite accuracy on some sets K_1, K_2 in infinite time. In such a case, we say that the identified $\widehat{f}$ has the *SEB property*. To our knowledge, no results exist on this SEB problem for the identified model.

In this note, we investigate the SEB problem using the Nonlinear Set Membership (NSM) method for the identification of nonlinear system developed in [9]. In that paper it is shown how to find an "optimal" estimate f_c of f_o from input–output data corrupted by bounded noise, not assuming a functional parametric form for f_o, but assuming a bound on f_o gradient norm. A validation result is given, used to derive "validated regions" within which to suitably choose the bounds on noise and on f_o gradient norm. A quantity r_I, called radius of information, is also derived, giving the worst-case L_p identification error $\|f_c(x_t, v_t) - f_o(x_t, v_t)\|_p$, for all $w_t = (x_t, v_t)$ in a domain of interest W. In this note, a particular case is considered of the so called local approach, which consists in considering a function f_a approximating f_o and on the application of the method described in [9] to the residue function $f_\Delta(w) \doteq f_o(w) - f_a(w)$. A linear approximation is considered here and, assuming that the approximating linear model is asymptotically stable (a necessary condition for obtaining the SEB property) a condition is given guaranteeing that, for all initial conditions x_o and input sequences v_t giving solutions of (1) in the domain W, the simulation error can be bounded as a function of r_I. The condition involves the bound on the residual function f_Δ gradient norm, so that selecting the bounds in the validated region and satisfying the previous condition, the resulting optimal NSM model f_c has the SEB property.

The note is organized as follows. In Section II, the NSM identification method is summarized. In Section III, sufficient conditions for NSM models stability and simulation error boundedness are given of the optimal NSM model f_c. In Section IV, a numerical example is shown in order to demonstrate the effectiveness of the presented theoretical results.

II. NSM Identification

In this section, the concepts and results of the NSM methodology developed in [9], needed for the development of the present note, are briefly recalled. Consider that a set of noise corrupted data $\widetilde{Y}_T = [\widetilde{y}_0, \widetilde{y}_1, \ldots, \widetilde{y}_T]$ and $\widetilde{W}_T = [\widetilde{w}_0, \widetilde{w}_1, \ldots, \widetilde{w}_T]$ generated by (1) is available. Then

$$\tilde{y}_{t+1} = f_o(\widetilde{w}_t) + e_t, \qquad t = 0, 1, \ldots, T-1$$

where the term e_t accounts for the fact y_{t+1} and w_t are not exactly known. The aim is to obtain an estimate $\widehat{f}$ of f_o, possibly giving small identification error $\|\widehat{f} - f_o\|_p$, where $\|\cdot\|_p$ is a L_p norm, $\|f\|_p \doteq [\int_W |f(w)|^p dw]^{1/p}, p \in [1, \infty), \|f\|_\infty \doteq \text{ess} - \sup_{w \in W} |f(w)|$ and W is a bounded subset of $\mathbb{R}^m$. It must be noted that no finite bound on the identification error can be guaranteed, unless some assumptions are made on the function f_o and the noise sequence e. 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 [10]. The typical approach in the literature is to assume a given parametric form for f_o (linear, bilinear, neural networks, etc.) and statistical models on the noise sequences. In the Nonlinear Set Membership approach presented in [9], the gradient of f_o and noise sequences are supposed to be bounded. Here, a particular case is considered of the so called local approach, which consists in the evaluation of a function f_a approximating f_o (using any desired method, e.g., PE method) and on the application of the method described in [9] to the residue function $f_\Delta(w) \doteq f_o(w) - f_a(w)$ with the assumption that $\|f'_\Delta(w)\| \leq \gamma$, where $f'_\Delta(w)$ denotes the gradient of $f_\Delta(w)$. This is equivalent to the locally varying (in the regressor space) bound $\|f'_a(w)\| - \gamma \leq \|f'_o(w)\| \leq \|f'_a(w)\| + \gamma$ for the gradient of function f_o. For a more detailed discussion on the advantages of using such a local NSM approach, see [9].

In this note we consider the case that f_a is a linear approximation of $f_o(w)$, i.e., $f_a(w) = \theta^T w$ (see, e.g., [11] and [12] for methods to estimate linear models of nonlinear systems) and we make the following assumptions.

1) Assumptions on $f_\Delta(w)$:

$$f_\Delta \in K^L \doteq \{g \in C^1(W), \|g'(w)\| \leq \gamma, \forall w \in W\}$$

■

2) Assumptions on Noise:

$$|e_t| \leq \varepsilon_t, \qquad t = 0, 1, \ldots, T.$$

■

A key role in Set Membership identification is played by the *Feasible Systems Set*, often indicated as "unfalsified functions set," i.e., the set of all functions consistent with prior information and measured data.

Definition 1: Feasible Systems Set

$$\text{FSS}_T \doteq \{f : f(w) = \theta^T w + g(w), g \in K^L, \ |\tilde{y}_{t+1} - f(\widetilde{w}_t)| \leq \varepsilon_t, t = 0, 1, \ldots, T-1\}.$$

■

The Feasible Systems Set FSS_T summarizes all the available information (measured data and prior information on f_o and noise e). If prior assumptions are "true," then $f_o \in \mathrm{FSS}_T$, an important property in view of subsequent use for prediction. 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: $\mathrm{FSS}_T \neq \emptyset$. ■

Conditions for prior assumptions validation are given as follows. Let us introduce the functions

$$\begin{aligned} \overline{f}_\Delta(w) &\doteq \min_{t=0,\dots,T-1}(\overline{h}_t + \gamma\|w - \widetilde{w}_t\|) \\ \underline{f}_\Delta(w) &\doteq \max_{t=0,\dots,T-1}(\underline{h}_t - \gamma\|w - \widetilde{w}_t\|) \\ \overline{h}_t &\doteq \widetilde{y}_{t+1} - \theta^T\widetilde{w}_t + \varepsilon_t \quad \underline{h}_t \doteq \widetilde{y}_{t+1} - \theta^T\widetilde{w}_t - \varepsilon_t. \end{aligned} \tag{3}$$

■

Result 1 [9]:

i) A necessary condition for prior assumptions to be validated is: $\overline{f}_\Delta(\widetilde{w}_t) \geq \underline{h}_t$, $t = 0, 1, \dots, T-1$.
ii) A sufficient condition for prior assumptions to be validated is: $\overline{f}_\Delta(\widetilde{w}_t) > \underline{h}_t$, $t = 0, 1, \dots, T-1$. ■

The previous result can be used for choosing values of the bounds $\varepsilon = [\varepsilon_0 \dots \varepsilon_T]$ and γ assuring that prior assumptions are not invalidated by data. The space (ε, γ) is divided in two regions. One region corresponds to values of ε and γ falsified by data ($\mathrm{FSS}_T = \emptyset$), the other corresponds to values of ε and γ validated by data ($\mathrm{FSS}_T \neq \emptyset$). In the space (ε, γ), the function $\gamma^*(\varepsilon) \doteq \inf_{\mathrm{FSS}_T \neq \emptyset} \gamma$ individuates a surface that separate falsified values of ε and γ from validated ones, Clearly, ε and γ must be chosen in the validated parameters region; see, e.g., Fig. 4 in the example section and [9] for a more detailed discussion and procedure for such a choice.

An identification algorithm ϕ is an operator mapping all available information about function f_o, noise e, data $(\widetilde{Y}_T, \widetilde{W}_T)$ until time T, summarized by FSS_T, into an estimate $\hat{f}$ of f_o : $\phi(\mathrm{FSS}_T) = \hat{f} \simeq f_o$. The related L_p error is: $e(\hat{f}) = e(\phi(\mathrm{FSS}_T)) = \|f_o - \hat{f}\|_p$. This error cannot be exactly computed, since it is only known that $f_o \in \mathrm{FSS}_T$, but its tightest bound is given by $e(\hat{f}) \leq \sup_{f \in \mathrm{FSS}_T} \|f - \hat{f}\|_p$. This motivates the following definition of the identification error, often indicated as local worst-case or guaranteed error,and of optimal algorithm.

Definition 3: The identification error of $\hat{f} = \phi(\mathrm{FSS}_T)$ is: $E(\hat{f}) \doteq \sup_{f \in \mathrm{FSS}_T} \|f - \hat{f}\|_p$. ■

Definition 4: An algorithm ϕ^* is called optimal if

$$E[\phi^*(\mathrm{FSS}_T)] \doteq \inf_{\phi} E[\phi(\mathrm{FSS}_T)] = \inf_{\widehat{f}} \sup_{f \in \mathrm{FSS}_T} \|f - \hat{f}\|_p = r_I.$$

■

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.

The following result shows that the algorithm $\phi_c(\mathrm{FSS}_T) = f_c$, where

$$f_c(w) \doteq \theta^T w + \frac{1}{2}[\underline{f}_\Delta(w) + \overline{f}_\Delta(w)] \tag{4}$$

and $\bar{f}_\Delta$ and $\bar{f}_\Delta$ are defined in (3), is optimal for any L_p norm.

Result 2 [9]: For any $L_p(W)$ norm, with $p \in [1, \infty]$

i) the identification algorithm $\phi_c(\mathrm{FSS}_T) = f_c$ is optimal;
ii) $E(f_c) = (1/2)\|\overline{f}_\Delta - \underline{f}_\Delta\|_p = r_I = \inf_\phi E[\phi(\mathrm{FSS}_T)]$.

III. Boundedness of the Simulation Error

In this section, conditions are derived, assuring that the model f_c identified by optimal algorithm ϕ_c has the SEB property, i.e., simulates f_o up to finite accuracy for any initial condition and input sequence in suitable domains and in infinite time. At first, a stability result for solutions of the optimal model f_c is given. Let

$$\Theta \doteq \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \cdots & \theta_{n-1} & \theta_n \\ 1 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{bmatrix} \in R^{n \times n}.$$

For matrix A, let $\|A\|$ denote its spectral norm. All the eigenvalues $\lambda_i(\Theta)$ of matrix Θ are strictly inside the unit circle if and only if the linear regression model $y_{t+1} = \theta^T w_t$ is asymptotically stable. Moreover, if $|\lambda_i(\Theta)| < 1, \forall i$, then the spectral norm of the powers of Θ is exponentially bounded, i.e., constants $L > 0$ and $\max_i |\lambda_i(\Theta)| \leq \rho < 1$ exist such that

$$\|\Theta^t\| \leq L\rho^t \qquad \forall t \geq 0. \tag{5}$$

The following theorem shows that if the linear approximation $\theta^T w$ of $f_o(w)$ is asymptotically stable and the residue function $f_\Delta(w) \doteq f_o(w) - \theta^T w$ has gradient not "too large," then the solutions of the optimal model f_c are uniformly exponentially stable.

Theorem 1: Let $\lambda_i(\Theta)$ be the eigenvalues of matrix Θ and $L > 0, \rho < 1$ constants such that (5) holds. Assume that

i) $|\lambda_i(\Theta)| < 1, \ \forall i$;
ii) $\gamma < (1 - \rho)/L$.

Then, for any initial condition x_o and input sequence v, the solution $y_t(f_c, x_0, v)$ is uniformly exponentially stable, i.e.,

$$|y_t(f_c, x_0', v) - y_t(f_c, x_0, v)| < L\widehat{\rho}^t\|x_0' - x_0\| \forall x_0', t$$

where

$$\widehat{\rho} = \rho + L\gamma < 1.$$

Proof: Consider the solutions $y_t(f_c, x_0', v)$ and $y_t(f_c, x_0, v)$ of system (1) with $f = f_c$, corresponding to the same input v and to two different initial conditions x_0' and x_0. Let

$$\begin{aligned} \delta y_{t+1} &\doteq y_{t+1}(f_c, x_0', v) - y_{t+1}(f_c, x_0, v) \\ &= f_c(x_t', v_t) - f_c(x_t, v_t) \end{aligned}$$

where x_t' and x_t are obtained by iteration of (1) with $f = f_c$, using input v and starting from initial conditions x_0' and x_0, respectively. Defining

$$f_{\Delta c}(w) \doteq \frac{1}{2}[\underline{f}_\Delta(w) + \overline{f}_\Delta(w)]$$

it follows

$$\delta y_{t+1} = f_c(x_t', v_t) - f_c(x_t, v_t) \tag{6}$$

$$= \theta^T(x_t' - x_t) + f_{\Delta c}(x_t', v_t) - f_{\Delta c}(x_t, v_t). \tag{7}$$

Now, suppose that $x'_t \neq x_t$ and consider the vector $a_t \in \mathbb{R}^n$

$$a_t \doteq \frac{f_{\Delta c}(x'_t, v_t) - f_{\Delta c}(x_t, v_t)}{(\|x'_t - x_t\|)^2} \times [(x'_t - x_t)_1 \ \cdots \ (x'_t - x_t)_n]^T \quad (8)$$

where $(x'_t - x_t)_i$ is the ith component of vector $(x'_t - x_t)$. Then

$$a_t^T(x'_t - x_t) = \frac{f_{\Delta c}(x'_t, v_t) - f_{\Delta c}(x_t, v_t)}{\|x'_t - x_t\|^2} \sum_{i=1}^{n} (x'_t - x_t)_i^2 = f_{\Delta c}(x'_t, v_t) - f_{\Delta c}(x_t, v_t).$$

Equation (6) can thus be written as

$$\delta y_{t+1} = (\theta + a_t)^T (x'_t - x_t) = (\theta + a_t)^T \delta x_t \quad (9)$$

where $\delta x_t \doteq (x'_t - x_t) = [\delta y_t \ \ldots \ \delta y_{t-n+1}]^T$. Equation (9) represents a linear time-variant system in regression form describing the time evolution of δy_{t+1}. Consider such system in the state–space form

$$\begin{aligned} \delta x_{t+1} &= (\Theta + A_t)\, \delta x_t \\ \delta y_t &= [1 \ 0 \ \cdots \ 0]\, \delta x_t \\ A_t &\doteq \begin{bmatrix} a_{t1} & a_{t2} & \cdots & a_{tn} \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{n \times n} \end{aligned} \quad (10)$$

where a_{ti} is the ith component of vector a_t. If all eigenvalues of matrix Θ are strictly inside the unit circle and if $\|A_t\| < (1-\rho)/L, \forall t$, a linear time-variant system of the form (10) is uniformly exponentially stable (see, e.g., [13, Th. 24.7]) and

$$\|\delta x_t\| \leq L(\rho + L \sup_{\tau \geq 0} \|A_\tau\|)^t \|\delta x_0\|, \qquad t = 0, 1, \ldots. \quad (11)$$

From the definition of matrix A_t, it follows:

$$\|A_t\| = \|a_t\| = \frac{|f_{\Delta c}(x'_t, v_t) - f_{\Delta c}(x_t, v_t)|}{\|x'_t - x_t\|^2} \times \sqrt{\sum_{i=1}^{n} (x'_t - x_t)_i^2} = \frac{|f_{\Delta c}(x'_t, v_t) - f_{\Delta c}(x_t, v_t)|}{\|x'_t - x_t\|}.$$

On the other hand, in [9] it is shown that function $f_{\Delta c}$ is Lipschitz continuous with Lipschitz constant γ, i.e.,

$$\frac{|f_{\Delta c}(x', v_t) - f_{\Delta c}(x, v_t)|}{\|x' - x\|} \leq \gamma \qquad \forall x', x, v_t.$$

Thus, $\max_\tau \|A_\tau\| \leq \gamma$ and, being $\gamma < (1-\rho)/L$ by assumption ii), from (11) it follows:

$$|y_t(f_c, x'_0, v) - y_t(f_c, x_0, v)| = |\delta y_t| \leq \|\delta x_t\| \leq L(\rho + L\gamma)^t \|\delta x_0\|, \qquad t = 0, 1, \ldots$$

and the claim is proved. ∎

Now, we are in position to prove the main result of the note, showing that the assumptions of Theorem 1 are sufficient conditions for the optimal model f_c to have the SEB property.

Theorem 2: Let $\lambda_i(\Theta)$ be the eigenvalues of matrix Θ and $L > 0, \rho < 1$ constants such that (5) holds. Assume that

i) $|\lambda_i(\Theta)| < 1, \forall i$;
ii) $\gamma < (1-\rho)/L$.

Then, for all initial conditions x_0 and inputs v giving solutions for f_o such that $(x_t, v_t) \in W \forall t$, a constant $K \in [0, \infty)$ exists such that the simulation error of the optimal estimate f_c is bounded as

$$|\mathrm{SE}_t(f_o, f_c)| \leq L\widehat{\rho}^t \|x_0 - \widetilde{x}_0\| + K r_I \quad \forall t$$
$$\text{where } \widehat{\rho} = \rho + L\gamma < 1 \text{ and } r_I = \frac{1}{2}\|\overline{f}_\Delta - \underline{f}_\Delta\|_\infty.$$

Proof: Let $\delta y_{t+1} \doteq y_{t+1}(f_c, \widetilde{x}_0, v) - y_{t+1}(f_o, x_0, v)$. Then

$$\begin{aligned} \delta y_{t+1} &= f_c(\widehat{x}_t, v_t) - f_o(x_t, v_t) \\ &= f_c(\widehat{x}_t, v_t) - f_c(x_t, v_t) \\ &\quad + f_c(x_t, v_t) - f_o(x_t, v_t) \end{aligned}$$

where $\widehat{x}_t$ is obtained by iteration of (1) with $f = f_c$, using input v and starting from initial conditions $\widetilde{x}_0$.

Using the same arguments as in the proof of Theorem 1 results in

$$\delta y_{t+1} = (\theta + a_t)^T [\delta y_t \ \ldots \ \delta y_{t-n+1}]^T + \xi_t \quad (12)$$

where a_t is given in (8) for $x'_t = \widehat{x}_t$ and $\xi_t = f_c(x_t, v_t) - f_o(x_t, v_t)$. In the proof of Theorem 1, it is also shown that the system $\delta y_{t+1} = (\theta + a_t)^T [\delta y_t \ \cdots \ \delta y_{t-n+1}]^T$ is uniformly exponentially stable. Then, in this case, δy_t is the solution of a uniformly exponentially stable linear system corresponding to initial condition $\widetilde{x}_0 - x_0 = [\widetilde{y}_0 - y_0 \ \cdots \ \widetilde{y}_{-n+1} - y_{-n+1}]^T$ and input $\xi_t = f_c(x_t, v_t) - f_o(x_t, v_t)$. Let us write the solution δy_t as the sum of the free-evolution part and the forced part

$$\delta y_t = \delta y_t^{\text{free}} + \delta y_t^{\text{forced}}.$$

From Theorem 1, we have that the free-evolution part is bounded as

$$\left|\delta y_t^{\text{free}}\right| < L\widehat{\rho}^t \|\widetilde{x}_0 - x_0\| \qquad \forall x'_0, t$$

where $\widehat{\rho} = \rho + L\gamma < 1$. On the other hand, the forced part is bounded as (see, e.g., [13, Lemma 27.4])

$$\left|\delta y_t^{\text{forced}}\right| \leq K \sup_{\tau \geq 0} |\xi_\tau| \qquad \forall t$$

for some $K \in [0, \infty)$. From assumption $(x_t, v_t) \in W \ \forall t$ and Result 2, it follows that

$$\sup_{\tau \geq 0} |\xi_\tau| = \sup_{\tau \geq 0} |f_c(x_\tau, v_\tau) - f_o(x_\tau, v_\tau)| \leq \sup_{w \in W} |f_c(w) - f_o(w)| = \|f_c - f_o\|_\infty.$$

Moreover, from Result 2, it follows that

$$\|f_c - f_o\|_\infty \leq r_I = \frac{1}{2}\|\overline{f}_\Delta - \underline{f}_\Delta\|_\infty.$$

Then, being $|\delta y_t| \leq |\delta y_t^{\text{free}}| + |\delta y_t^{\text{forced}}|$ and $\mathrm{SE}_t = \delta y_t$, the claim follows. ∎

Remark 1: Theorem 2 gives sufficient conditions for optimal model f_c to have the SEB property. As it happens when sufficient conditions are derived, the problem of their conservativeness arises. Indeed, condition i) is also necessary, as evident from the fact that if $f_o(w) = \theta_o^T w$ is linear, even in case that an exact linear approximation $f_a(w) = \theta_o^T w$ is used, its asymptotic stability is necessary for f_c to have the SEB property. As regard to condition ii), the example presented in the next section shows that it can be actually met even in cases of strong nonlinearities, such as saturation.

Remark 2: Theorem 2 gives guidelines for the choice of the values of the bounds γ and ε, additional to the ones given in [9]. As recalled in Section II, the values of bounds γ and ε have to be chosen in the

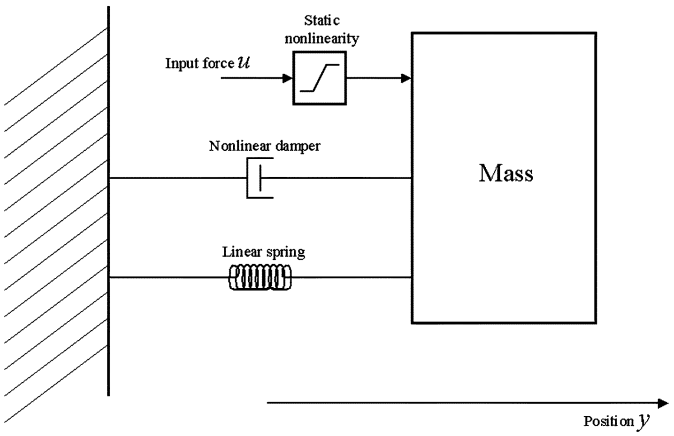

Fig. 1. Nonlinear mass-spring-damper system.

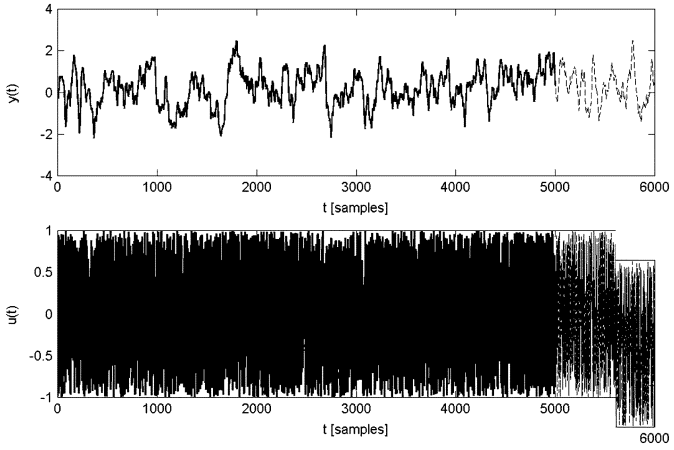

Fig. 2. Estimation data set (bold line) and validation data set (dashed line).

validated region which can be computed using the validation Result 1; see, e.g., Fig. 4 in the example section. Theorem 2 suggests to choose, if possible, a value of $\gamma < (1-\rho)/L$ in such a region.

IV. Example

A set of 6000 data has been generated from the following nonlinear system:

$$y_{t+1} = 1.8y_t - 0.82y_{t-1} + 0.0024\sin(y_{t-1}) + 0.047\tanh(3u_t)$$

representing a discrete-time approximation of a mass-spring-damper system with linear spring, nonlinear damper and input saturation (see Fig. 1). Input u is the force acting on the mass and output y is the mass position.

A uniform i.i.d. random input of amplitude ≤ 1 has been used. The output data of the estimation set have been corrupted by a uniform i.i.d. random additive noise of amplitude ≤ 0.025. The set of data is shown in Fig. 2. The first 5000 data, called estimation set, have been used for model identification, the remaining 1000 data, called validation set, have been used for model testing.

All the models have been identified using as regressor $w_t = [y_t \;\; y_{t-1} \;\; u_t]^T$.

Linear Output Error (OE) Model: The OE model is

$$f(w) = \theta^T w$$

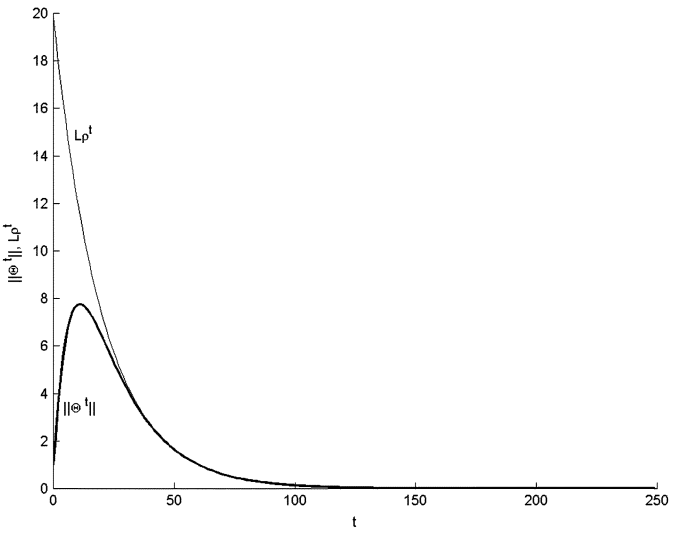

Fig. 3. $||\Theta^t||$ (bold line) and $L\rho^t$ (thin line) sequences.

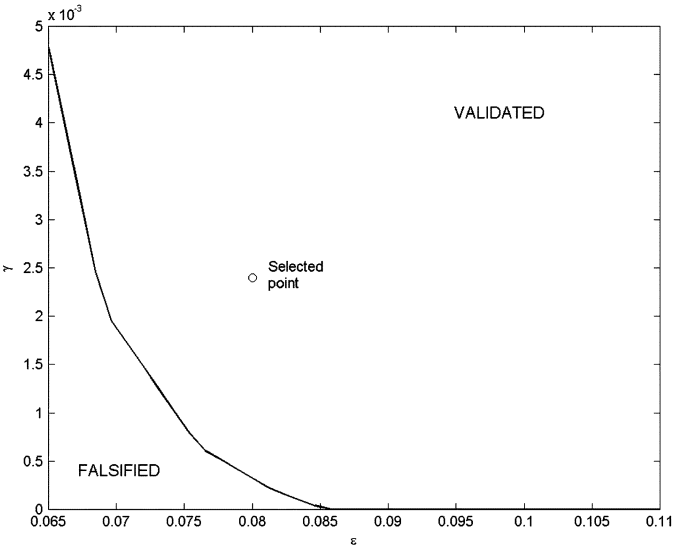

Fig. 4. Validation regions for model NSM.

where $\theta = [1.8 \;\; -0.81 \;\; 0.06]^T$ has been estimated by means of the Matlab Systems Identification Toolbox using the output error estimation method. This model is asymptotically stable.

NSM Model: The NSM model is:

$$f_c(w) = \theta^T w + \frac{1}{2}[\underline{f}_\Delta(w) + \overline{f}_\Delta(w)]$$

where θ is the one of the previous OE model.

The values of L and ρ for which $\|\Theta^t\| \leq L\rho^t \; \forall t$ are taken as $L = 19.8$ and $\rho = 0.952$. In Fig. 3, the sequences $\|\Theta^t\|$ and $L\rho^t$ are shown.

The functions $\underline{f}_\Delta(w)$ and $\overline{f}_\Delta(w)$ have been evaluated as given in Section II, choosing the following values of γ and $\varepsilon_t = \varepsilon, \forall t$ in the validated region and satisfying the assumption of Theorem 2: $\gamma = 0.0024$ $\varepsilon = 0.08$. Note that $(1-\rho)/L = 0.002\,44$. The validation regions for model NSM are shown in Fig. 4

Neural Network Models $\mathrm{NN}_{\mathrm{narx}}$ *and* $\mathrm{NN}_{\mathrm{noe}}$*:* The $\mathrm{NN}_{\mathrm{narx}}$ and $\mathrm{NN}_{\mathrm{noe}}$ models have been obtained considering a two-layer neural network (see, e.g., [2]–[5]) for the regression function:

$$f(w) = \sum_{i=1}^{r} \alpha_i \sigma\left(\beta_i^T w - \lambda_i\right) + \zeta.$$

TABLE I
ONE-STEP AHEAD PREDICTION AND SIMULATION ERRORS

Model	NSM	OE	$\mathrm{NN_{narx}}$	$\mathrm{NN_{noe}}$
$\mathrm{RMSE_P}$	0.005	0.011	0.008	0.009
$\mathrm{RMSE_S}$	0.091	0.267	0.299	0.262

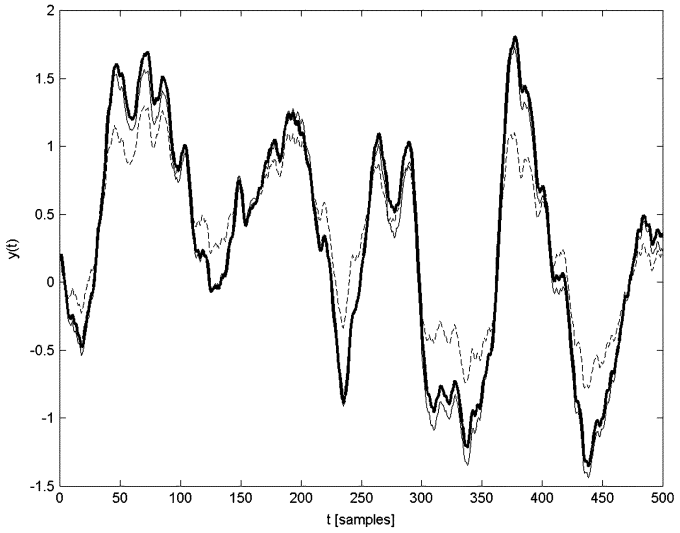

Fig. 5. Validation set: data (bold line), NSM (thin line), and NNnoe simulation (dashed line).

Here, $\alpha_i, \lambda_i, \zeta \in \mathbb{R}, \beta_i \in \mathbb{R}^n$ are parameters and $\sigma(x) = 2/(1 + e^{-2x}) - 1$ is a sigmoidal function. Several NARX and NOE neural network models with different values of r (from $r = 3$ to $r = 16$) have been trained on the estimation set using the Matlab Neural Networks Toolbox. Note that NOE neural networks models cannot be directly estimated using the MATLAB Neural Network Toolbox. Indeed, it is required to set the neurons connections in order to obtain a suitable recursive network.

The NARX model with $r = 8$ showing the best simulation performances, has been taken for model $\mathrm{NN_{narx}}$. 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 $\mathrm{NN_{narx}}$ model. This NOE model, showing a slight improvement in simulation performances with respect to the $\mathrm{NN_{narx}}$ one, has been taken for model $\mathrm{NN_{noe}}$. ■

In Table I, the root mean square errors obtained by the identified models on the validation data set are reported. The simulation error is indicated as $\mathrm{RMSE_S}$, the one-step ahead prediction error is indicated as $\mathrm{RMSE_P}$. In Fig. 5, a portion of validation data and NSM and $\mathrm{NN_{noe}}$ models simulation are shown.

It can be noted that the accuracy improvements of the NSM model over the $\mathrm{NN_{narx}}$ and $\mathrm{NN_{noe}}$ models, though moderate for one-step ahead prediction, are quite significant in simulation.

V. CONCLUSION

In this note, a first result is derived on the quality of identified models of nonlinear systems measured by the accuracy in simulating the system behavior for future inputs not used in the identification. Models identified by classical methods minimizing the prediction error, do not necessary give "good" simulation error on future inputs and even the SEB property is not guaranteed. The NSM approach proposed in [9] is taken, assuming that the nonlinear regression function, representing the difference between the system to be identified and a linear approximation, has gradient norm bounded by a constant γ. Moreover, the noise sequence is assumed unknown but bounded.

If the chosen linear approximation is asymptotically stable (a necessary condition for the SEB property), a condition on γ is given, guaranteeing that the simulation error is bounded by a function of the radius of information r_I, so that the identified NSM model f_c has the SEB property. The bounding constants γ and ε, defining the SM assumptions on system and noise, have to be selected in the validated region, i.e., in such a way that assumptions are consistent with measured data. Thus, if values in this region exist, satisfying the sufficient condition, the results derived in note can be used to give guidelines for the choice of the bounding constants γ and ε, additional to the ones in [9] and useful for obtaining models with "low" simulation errors.

The numerical example, representing a mass-spring-damper system with nonlinear damper and input saturation, demonstrates the effectiveness of the presented approach. Indeed, the $\mathrm{NN_{narx}}$ and $\mathrm{NN_{noe}}$ models identified in this example have similar one-step ahead errors as the NSM model. However, the NSM model, satisfying the derived sufficient conditions, displays a simulation error significantly lower (more than three times) than the simulation errors obtained by the $\mathrm{NN_{NARX}}$ and $\mathrm{NN_{noe}}$ models.

Many important open problems remain, deserving further investigations, such as giving conditions on input and noise assuring asymptotic convergence to zero of the radius of information, evaluation of constant k in Theorem 2, derivation of identification algorithms directly minimizing the simulation error.

REFERENCES

[1] R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic systems—A survey on input/output approaches," *Automatica*, vol. 26, pp. 651–677, 1990.
[2] 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.
[3] K. S. Narendra and S. Mukhopadhyay, "Neural networks for system identification," *Sysid 97*, vol. 2, pp. 763–770, 1997.
[4] R. Isermann, S. Ernst, and O. Nelles, "Identification with dynamic neural networks -architectures, comparisons, applications," *Sysid 97*, vol. 3, pp. 997–1022, 1997.
[5] M. Norgaard, O. Ravn, N. Poulsen, and L. K. Hansen, *Neural Networks for Modeling and Control of Dynamic Systems*. New York: Springer-Verlag, 2000.
[6] E. D. Sontag, "Neural nets as systems models and controllers," in *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, 1992, pp. 73–79.
[7] E. B. Kosmatopoulos, M. M. Polycarpou, M. A. Christodoulou, and P. A. Ioannou, "High-order neural network structures for identification of dynamical systems," *IEEE Trans. Autom. Control*, vol. 6, no. 2, pp. 442–431, Feb. 1995.
[8] M. B. Matthews and G. S. Moschytz, "The identification of nonlinear discrete-time fading-memory systems using neural networks models," *IEEE Trans. Autom. Control*, vol. 40, no. 11, pp. 740–751, Nov. 1994.
[9] M. Milanese and C. Novara, "Set membership identification of nonlinear systems," *Automatica*, vol. 40/6, pp. 957–975, 2004.
[10] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC: Winston, 1977.
[11] P. Mäkilä and J. Partington, "On linear models for nonlinear systems," *Automatica*, vol. 39, pp. 1–13, 2003.
[12] L. Ljung, "Estimating linear time-invariant models of nonlinear time-varying systems," *Eur. J. Control*, vol. 7, pp. 203–219, 2003.
[13] W. J. Rugh, *Linear System Theory*. Upper Saddle River, NJ: Prentice-Hall, 1996.