Exponentially Weighted Least Squares Identification of Time-Varying Systems with White Disturbances

Marco C. Campi

Abstract— **This paper is devoted to the stochastic analysis of recursive least squares (RLS) identification algorithms with an exponential forgetting factor. A persistent excitation assumption of a conditional type is made that does not prevent the regressors from being a dependent sequence. Moreover, the system parameter is modeled as the output of a random-walk type equation without extra constraints on its variance. It is shown that the estimation error can be split into two terms, depending on the parameter drift and the disturbance noise, respectively. The first term turns out to be proportional to the memory length of the algorithm, whereas the second is proportional to the inverse of the same quantity. Even though these dependence laws are well known in very special mathematical frameworks (deterministic excitation and/or independent observations), this is believed to be the first contribution where they are proven in a general dependent context. Some idealized examples are introduced in the paper to clarify the link between generality of assumptions and applicability of results in the developed analysis.**

I. Introduction and Preliminaries

A. The Exponentially Weighted Least Squares Algorithm

GIVEN a sequence of stochastic observation vectors $\varphi(\cdot) \in R^n$, consider the scalar process $y(\cdot)$ generated according to the following time-varying equation

$$y(t) = \varphi(t)'\vartheta^\circ(t) + d(t). \tag{1a}$$

In (1a), the scalar $d(\cdot)$ is a disturbance term, and $\vartheta^\circ(\cdot) \in R^n$ is a stochastic sequence of unknown parameter vectors, whose time evolution satisfies a random-walk type equation:

$$\vartheta^\circ(t+1) = \vartheta^\circ(t) + \delta\vartheta^\circ(t). \tag{1b}$$

The generality of model (1) can be appreciated by means of the following examples, which will be used later on to test the applicability of the theory developed in the present paper.

Examples:

1) Linear combiner [1]: The linear combiner is an adaptive nonrecursive filter widely used in adaptive signal processing, whose general form is

$$y(t) = a_1(t)u_1(t) + a_2(t)u_2(t) + \cdots + a_n(t)u_n(t) + d(t)$$

where $u_i(t)$ and $i = 1, 2, \ldots, n$ are exogenous inputs. The linear combiner takes the form (1a) by introducing the observation vector

$$\varphi(t) = [u_1(t) \quad u_2(t) \cdots u_n(t)]'.$$

2) Autoregressive model [2]: Autoregressive models are often used in time series analysis since they provide parsimonious representations for both slow and fast dynamics hidden in the data. The time-varying version is

$$y(t) = a_1(t)y(t-1) + a_2(t)y(t-2) + \cdots + a_n(t)y(t-n) + d(t)$$

which can be given form (1a) with the notation

$$\varphi(t) = [y(t-1)\ y(t-2) \cdots y(t-n)]'.$$

3) Hammerstein model [3]: When the input/output dynamics of a system contains a nonlinear but fast component, a common modeling procedure consists of resorting to a linear dynamic model complemented with a nonlinear gain. Then, the nonlinear gain is represented by a Taylor expansion as follows (Hammerstein model):

$$y(t) = a_0(t) + a_1(t)u(t) + a_2(t)u(t)^2 + \cdots + a_n(t)u(t)^n + d(t).$$

In addition, this model can be put in form (1a) by introducing the vector

$$\varphi(t) = [1\ u(t)\ u(t)^2 \cdots u(t)^n]'.$$

□

The following assumptions on $d(\cdot)$ and $\delta\vartheta^\circ(\cdot)$ will be assumed to hold throughout the paper (the symbol $\sigma(v)$, where v is a set of random variables, stands for the σ-algebra generated by v).

A.1 $d(t)$ is a zero expected value random variable independent of $\sigma(\varphi(i), \delta\vartheta^\circ(i), d(i-1), i \le t)$ and $E[d(t)^2] = \sigma^2$.

A.2 $\delta\vartheta^\circ(t)$ is a zero expected value random variable independent of $\sigma(\varphi(i), d(i), \delta\vartheta^\circ(i-1), i \le t)$ and $E[\|\delta\vartheta^\circ(t)\|^2] = \Delta^2$.

Manuscript received May 5, 1992; revised February 14, 1994. This work was supported by the Ministry of University and Technical and Scientific Research (MURST) under the projects "Model Identification, Systems Control and Signal Processing" and "Adaptive and Robust Control of Dynamical Systems." The associate editor coordinating the review of this paper and approving it for publication was Dr. H. Fan.

The author is with the Dipartimento di Elettronica per l'Automazione, Universita' di Brescia, Brescia, Italy.

IEEE Log Number 9404794.

For the estimation of the unknown parameters $\vartheta^\circ(t)$, we will consider the exponentially weighted least squares (EWLS) algorithm obtained by minimizing the loss function:

$$J(\vartheta) = \sum_{\tau=-\infty}^{t} \mu^{t-\tau}[\bar{y}(\tau) - \bar{\varphi}(\tau)'\vartheta]^2 \tag{2}$$

where

$$\bar{\varphi}(\tau) = \begin{cases} \varphi(\tau), & \text{if } \|\varphi(\tau)\| \le b_\varphi \\ \frac{b_\varphi}{\|\varphi(\tau)\|}\varphi(\tau), & \text{if } \|\varphi(\tau)\| > b_\varphi \end{cases} \tag{3a}$$

$$\bar{y}(\tau) = \begin{cases} y(\tau), & \text{if } \|\varphi(\tau)\| \le b_\varphi \\ \frac{b_\varphi}{\|\varphi(\tau)\|}y(\tau), & \text{if } \|\varphi(\tau)\| > b_\varphi. \end{cases} \tag{3b}$$

In (2), the coefficient $\mu \in (0,1)$ is the so-called *forgetting factor*, which is introduced in the loss function to discount old data in favor of fresh information. The lower the forgetting factor, the higher the algorithm responsiveness. The selection of the value for μ is a user's choice and is discussed, e.g., in [4] and [5]; see also [6] and [7]. In the sequel, the time constant $\lambda = 1/(1-\mu)$ associated with the discrete exponential function μ^t will be called the *memory length* of the algorithm.

The data $\varphi(\cdot)$ and $y(\cdot)$ are "cut" according to (3a) and (3b) before they are processed by the identification algorithm. Such a procedure avoids that data with large value due to oversized shots of noise or measurements errors have an excessive weight in the loss function (2). In this connection, the positive constant b_φ should be chosen in the light of *a priori* knowledge about the normal range of variability of the $\varphi(\cdot)$ entries; see [8] for more discussion.

The minimum of the loss function (2) is achieved by the vector ([8])

$$\hat{\vartheta}(t) = P(t)R(t) \tag{4a}$$

where

$$R(t) = \sum_{\tau=-\infty}^{t} \mu^{t-\tau}\bar{\varphi}(\tau)\bar{y}(\tau), \tag{4b}$$

and

$$P(t) = \left(\sum_{\tau=-\infty}^{t} \mu^{t-\tau}\bar{\varphi}(\tau)\bar{\varphi}(\tau)'\right)^{-1} \tag{4c}$$

is the so-called *covariance matrix* of the algorithm.

Equation (4) is the *batch* version of EWLS. It is well known that the algorithm can be given *recursive* form suitable for on-line identification. Numerically robust versions based on U-D factorization can be found, e.g., in [4]; in addition, see [9].

The present work aims at studying the performance of the EWLS algorithm in a stochastic framework. In Section I-B, we illustrate the basic problems arising in the analysis of least squares techniques; moreover, we will present our main results, putting them into perspective within the existing literature on the subject.

B. A Glimpse through the Existing Literature on Adaptive RLS Techniques

The following two questions motivate almost all the papers pertaining to the performance analysis of adaptive identification algorithms:

i) Is the algorithm able to keep the estimation error bounded—in some sense—in a range of situations sufficiently large with respect to its normal area of application?

ii) What does the estimation error depend on and in what way?

As for this last question, a key issue is finding the dependence of the estimantion error on the "tuning knobs" of the algorithm. This issue has been recently investigated in [10] with regard to the least mean squares (LMS) algorithm (see also [11]), where it is shown that the mean square of the estimation error can be given the following bound: (α = algorithm stepsize):

$$E[\|\hat{\vartheta}(t) - \vartheta^\circ(t)\|^2] \le c'\alpha + c''\frac{1}{\alpha}. \tag{5}$$

The two terms on the right-hand side of (5) describe how the noise and the parameter drift affect the estimation error. Note that the *steady-state error* increases linearly with α and the *tracking error* is inversely proportional to such a quantity.

Turning now to least squares techniques, to better describe the relevant results in the literature, we are well advised to consider the following general weighted least squares (WLS) algorithm:

$$\hat{\vartheta}(t) = \left(\sum_{\tau=-\infty}^{t} w(t,\tau)\ \psi(\tau)\psi(\tau)'\right)^{-1} \times \left(\sum_{\tau=-\infty}^{t} w(t,\tau)\psi(\tau)v(\tau)\right). \tag{6}$$

In (6), $\psi(\cdot)$ and $v(\cdot)$ are the "filtered observation vector" and "filtered output" obtained from $\varphi(\cdot)$ and $y(\cdot)$ via some processing such as prefiltering, cutting etc. Note that algorithm (4) corresponds to (6) with $\psi(\cdot) = \bar{\varphi}(\cdot)$, $v(\cdot) = \bar{y}(\cdot)$, and $w(t,\tau) = \mu^{t-\tau}$ as weights. In contrast with LMS, the WLS algorithm is characterized by the auxiliary memory constituted by the matrix $A(t) = \sum_{\tau=-\infty}^{t} w(t,\tau)\ \psi(\tau)\psi(\tau)'$ (information matrix). Loosely speaking, this matrix keeps memory of the total amount of information available in the different directions in the parameter space. As a consequence, the estimation at time t depends in a complex way on the entire history of data.

The first study of least squares techniques worth mentioning can be found in [12]. In this paper, the analysis relies on the independence assumption of regressors and the stationarity of observations. Even though based on these strong assumptions, this paper is illuminating in that it shed light for the first time on how the performance of the algorithm depends on its memory length. Moreover, due to averaging effects, the analysis carried out under the independence assumption is able, to some extent, to capture the basic behaviors exhibited by the algorithm even in the dependent context in the case of long

memory length algorithms; see [13] for another study in a similar mathematical context.

In [12]—as in almost all papers on the subject—the major technical difficulties arise for the possible lack of information when the observation vectors are described in a stochastic way. This is due to the fact that the sophisticated treatment of data performed by WLS algorithms is quite involved in the way in which uncertain information can possibly compensate for errors in the estimate. Obviously, a drastic simplification is obtained if one assumes that the observation vectors satisfy a stiff deterministic constraint of the type

$$cI \geq \sum_{i=\tau+1}^{\tau+s} \varphi(i)\varphi(i)' \geq cI, \forall \tau. \tag{7}$$

Basically, this condition imposes that a certain degree of information is available in any direction of the parameter space over any time interval of length s. In the analysis, this enables one to disregard the actual pattern of information, taking into account just the lower bound guaranteed by (7). Under condition (7), nice bounds for the estimation error have been worked out in [14] and [15].

More recently, Niedzwiecki and Guo [16] observed that under stationarity assumption on the observation vectors, the auxiliary matrix $A(t)^{-1} = \left(\sum_{\tau=-\infty}^{t} w(t,\tau)\psi(\tau)\psi(\tau)'\right)^{-1}$ tends to $B^{-1} = \sum_{\tau=-\infty}^{t} (w(t,\tau)E[\psi(\tau)\psi(\tau)'])^{-1}$ when the memory length of the algorithm tends to infinite. On the grounds of this consideration, they proposed to approximate the WLS estimate $\hat{\vartheta}(t)$ with the following "idealized" WLS estimate:

$$\hat{\hat{\vartheta}}(t) = \left(\sum_{\tau=-\infty}^{t} w(t,\tau)E[\psi(\tau)\psi(\tau)']\right)^{-1} \times \left(\sum_{\tau=-\infty}^{t} w(t,\tau)\psi(\tau)v(\tau)\right).$$

In this way, one has to deal with the constant matrix B^{-1} instead of the more complicated information matrix $A(t)^{-1}$. Applying this idea to the rectangular window algorithm (i.e., $w(t,\tau) = 1/N, t-\tau \leq N-1$; $w(t,\tau) = 0, t-\tau \geq N$), the authors were able to prove that for Gaussian regressors, the L^2 norm of the estimation error can be given the bound

$$E[\|\hat{\vartheta}(t) - \vartheta^{\circ}(t)\|^2] \leq c'N + c''\frac{1}{N}. \tag{8}$$

Note that (8) can be seen as the counterpart of (5) for the rectangular window LS algorithm. In analogy with the derivation of (5) given in [10], result (8) is also based on independence assumptions between $\varphi(\cdot)$, $\vartheta^{\circ}(\cdot)$, and $d(\cdot)$. Moreover, as already noted, the approach proposed in [15] is inherently based on stationarity assumption on the observation vector sequence $\varphi(\cdot)$. As such, it is suitable for linear combiner-type models only.

To the best knowledge of the present author, the first paper where WLS algorithms were studied without resorting to mutual independence among observation vector, noise, and parameter drift and/or stationarity assumption on the sequence $\varphi(\cdot)$ is [17]. There, a general algorithm with time-varying forgetting factor is considered, which is obtained by setting $w(t,\tau) = \pi_{j=0}^{t-\tau}\mu(t-j)$ in (6). In [17], only the first question set at the beginning of this section has been addressed. The main contribution consists of recognizing that the L^2 boundedness of the estimation error is strictly related to the L^1-boundedness of the so-called covariance matrix of the algorithm. The interested reader is also referred to [18] and [19] for a preliminary study in the case of ideal systems (no measurement noise and time-invariant true parameterization) carried out in the same mathematical framework as [17].

In this paper, we are basically concerned with the second question posed at the beginning of this section. More precisely, we want to work out bounds for the estimation error of the EWLS algorithm with the objective of clarifying how the performance of this algorithm depends on its memory length.

The main results of the paper can be summarized as follows:

i) The L^2 norm of the estimation error can be split into two terms, respectively, depending on the parameter drift and the disturbance. Under suitable persistent excitation conditions, the first term is proportional to the memory length of the algorithm, and the second one is inversely proportional to the same quantity.

ii) When $\Delta = 0$ (no drift in the parameter), the L^2 norm of the estimation error tends to zero if the forgetting factor μ tends to 1.

The results stated above have been worked out in a very general stochastic framework. In particular, we do not require any independence assumption of the observation vectors. Even though the independence assumption does not reflect the reality, its use has been common practice in the study of adaptive filtering algorithms. In this paper, we have been able to drop this assumption thanks to a novel approach based on the introduction of a "fake information matrix," which keeps memory of the "independent portion of information" (see Section IV for details). It is the hope (and the belief) of the present author that such a new approach can be helpful for the study of many more situations of interest besides that treated in this paper.

II. A Fundamental Expression for the Estimation Error

Let

$$\bar{d}(t) = \begin{cases} d(t), & \text{if } \|\varphi(t)\| \leq b_{\varphi} \\ \frac{b_{\varphi}}{\|\varphi(t)\|}d(t), & \text{if } \|\varphi(t)\| > b_{\varphi}. \end{cases}$$

The EWLS estimation $\hat{\vartheta}(t)$ given by (4) can be handled as follows:

$$\hat{\vartheta}(t) = \sum_{\tau=-\infty}^{t} \mu^{t-\tau}P(t)\bar{\varphi}(\tau)\left[\bar{\varphi}(\tau)'\vartheta^{\circ}(\tau) + \bar{d}(\tau)\right]$$

(using (1a))

$$= \sum_{\tau=-\infty}^{t} \mu^{t-\tau}P(t)\left[P(\tau)^{-1} - \mu P(\tau-1)^{-1}\right] \times \left[\vartheta^{\circ}(\tau+1) - \delta\vartheta^{\circ}(\tau)\right]$$

$$+\sum_{\tau=-\infty}^{t}\mu^{t-\tau}P(t)\bar{\varphi}(\tau)\bar{d}(\tau) \qquad \text{(using (1.b))}$$

$$=\vartheta^{\circ}(t+1)-\sum_{\tau=-\infty}^{t}\mu^{t-\tau}P(t)P(\tau)^{-1}\delta\vartheta^{\circ}(\tau)$$

$$+\sum_{\tau=-\infty}^{t}\mu^{t-\tau+1}P(t)P(\tau-1)^{-1}$$

$$\times[\vartheta^{\circ}(\tau)+\delta\vartheta^{\circ}(\tau)-\vartheta^{\circ}(\tau+1)]$$

$$+\sum_{\tau=-\infty}^{t}\mu^{t-\tau}P(t)\bar{\varphi}(\tau)\bar{d}(\tau)$$

$$=\vartheta^{\circ}(t+1)-\sum_{\tau=-\infty}^{t}\mu^{t-\tau}P(t)P(\tau)^{-1}\delta\vartheta^{\circ}(\tau)$$

$$+\sum_{\tau=-\infty}^{t}\mu^{t-\tau}P(t)\bar{\varphi}(\tau)\bar{d}(\tau) \qquad \text{(using (1b)).}$$

Consequently, the estimation error $\tilde{\vartheta}(t)=\hat{\vartheta}(t)-\vartheta^{\circ}(t+1)$ is given by

$$\tilde{\vartheta}(t)=\sum_{\tau=-\infty}^{t}\mu^{t-\tau}P(t)\left[-P(\tau)^{-1}\delta\vartheta^{\circ}(\tau)+\bar{\varphi}(\tau)\bar{d}(\tau)\right]. \quad (9)$$

Equation (9) constitutes the starting point for the analysis of the tracking properties of the EWLS algorithm, which is worked out in the next sections.

III. A Simple Independent Case

Assumptions A.1 and A.2 of Section I-A require the observation vector sequence $\varphi(i), i\le t$ to be independent of $d(t)$ and $\delta\vartheta^{\circ}(t)$, that is, the disturbance acting on the system and the variation in the true parameterization occurring at time t are unpredictable on the basis of the observation vectors measured up to time t. This working assumption looks very mild and is generally satisfied even if the observation vector contains an autoregressive part. However, the study of algorithm (4) under the only independence assumptions A.1 and A.2 constitutes, in general, a hard task. A considerable simplification in the analysis is achieved by introducing the following stronger independence assumption:

A.3 $\varphi(\cdot)$ is a sequence of independent variables such that $\sigma(\varphi(\cdot))$ is independent of $\sigma(d(\cdot),\delta\vartheta^{\circ}(\cdot))$.

Note that A.3 is a reasonable assumption as long as the observation vector does not contain past values of the system output, and the system inputs are not computed from a feedback law. This is, for instance, the case of Examples 1 and 3 of Section I-A.

For this section, we also introduce the following condition:

A.4 $\exists s,k>0$ such that

$$E\left[\left(\lambda_{\min}\left\{\sum_{i=\tau+1}^{\tau+s}\bar{\varphi}(i)\bar{\varphi}(i)'\right\}\right)^{-2}\right]\le k, \qquad \forall\tau.$$

This assumption is a persistent excitation condition on data. Apparently, for this condition to be satisfied, the span of the observation vectors over any time interval of length s has to be the entire parameter space with probability one. Even more so, A.4 imposes that the events on which information is poor in some direction are small enough so that $\lambda_{\min}\{\sum_{i=\tau+1}^{\tau+s}\bar{\varphi}(i)\bar{\varphi}(i)'\}$ is invertible in the mean squared sense.

Assumption A.4 constitutes an implicit condition on observation vectors. Obviously, it is interesting to work out explicit conditions on $\varphi(\cdot)$, guaranteeing that A.4 is met. In this connection, note that a strictly related problem has been recently discussed in [16]. To be precise, in this paper, the authors investigate the boundedness of the matrix $E[(\sum_{i=0}^{N-1}\psi(i)\psi(i)')^{-1}]$, where $\psi(\cdot)$ is an i.i.d. sequence with finite second-order moment. They show that the boundedness property is satisfied if and only if the following condition holds true (see [16] for the interpretation of this condition):

$$\exists\gamma>0,\delta>0,x_0>0 \text{ such that } \sup_{\|\beta\|=1}\left((p\beta'\psi(t))^2<x\right)\le\gamma x^{\delta},\forall x:0<x<x_0,\forall t. \quad (10)$$

Going through the proof of this result and taking into account the independence property of vectors $\psi(\cdot)$, it is not difficult to see that the existence of an integer N such that matrix $\sum_{i=0}^{N-1}\psi(i)\psi(i)'$ is invertible in the mean sense is equivalent to the existence of an integer M such that $\sum_{i=0}^{M-1}\psi(i)\psi(i)'$ is invertible in the mean squared sense. Starting from this consideration, some simple elaborations of the rationale in [16] allow one to conclude that condition (10)—with $\bar{\varphi}(t)$ in place of $\psi(t)$—is also necessary and sufficient for our assumption A.4 to be satisfied.

Equation (9) points out the major role played by matrix $P(t)$ in determining the amplitude of the estimation error $\tilde{\vartheta}(t)$. The proposition below provides a quantitative upper bound for the size of this matrix when assumptions A.3 and A.4 are met.

Proposition: Under assumptions A.3 and A.4, $\limsup_{\mu\to1}\|P(t)\|_{L^2}/(1-\mu)\le sk^{1/2}$.

Remark: Note that the dependence law of matrix $P(t)$ on the value of the forgetting factor μ coincides with that valid under deterministic excitation assumptions; see, e.g., [15].

Proof: Let

$$\Phi(t_1,t_2)=\sum_{i=t_1}^{t_2}\mu^{t_2-i}\bar{\varphi}(i)\bar{\varphi}(i)'.$$

The following recursive inequality for $\lambda_{\min}\{\Phi(-\infty,\cdot)\}$ holds true (see the Appendix for its derivation):

$$\left\{E\left[(\lambda_{\min}\{\Phi(-\infty,\tau)\})^{-2}\right]\right\}^{-1/2}\ge\left\{E\left[(\lambda_{\min}\{\Phi(\tau-s+1,\tau)\})^{-2}\right]\right\}^{-1/2}+\mu^{s}\left\{E\left[(\lambda_{\min}\{\Phi(-\infty,\tau-s)\})^{-2}\right]\right\}^{-1/2},\forall\tau. \quad (11)$$

By iteratively using inequality (11) starting from $\tau = t$ one gets

$$\left\{E\left[(\lambda_{\min}\{\Phi(-\infty,t)\})^{-2}\right]\right\}^{-1/2} \geq \sum_{\tau=0}^{\infty}\mu^{\tau s}\{E[(\lambda_{\min}\{\Phi(t-(\tau+1)s+1, t-\tau s)\})^{-2}]\}^{-1/2}. \quad (12)$$

The statement of the proposition easily follows by observing that the left-hand side of (12) coincides with $(\|P(t)\|_{L^2})^{-1}$ and that, thanks to Assumption A.4, $\{E[(\lambda_{\min}\{\Phi(t-(\tau+1)s+1, t-\tau s)\})^{-2}]\}^{-1/2} \geq \mu^{s-1}k^{-1/2}, \forall\tau$. □

Now, consider the inverse of the covariance matrix $(P(t)^{-1})$. Observing that vectors $\bar{\varphi}$ are bounded from above by b_φ (see (3a)), from definition (4c) of matrix $P(t)$, it follows that $\|P(t)^{-1}\| \leq b_\varphi^2/(1-\mu), \forall t$. Hence, bearing in mind the independence Assumption A.3, expression (9) for the estimation error gives

$$\begin{aligned}\|\tilde{\vartheta}(t)\|_{L^2}^2 &= \sum_{\tau=-\infty}^{t}\mu^{2(t-\tau)}\left(E\left[\|P(t)P(\tau)^{-1}\delta\vartheta^\circ(\tau)\|^2\right] + E\left[\|P(t)\bar{\varphi}(\tau)\bar{d}(\tau)\|^2\right]\right)\\ &\leq \sum_{\tau=-\infty}^{t}\mu^{2(t-\tau)}\left(E[\|P(t)\|^2[b_\varphi^4/(1-\mu)^2]\|\delta\vartheta^\circ(\tau)\|^2] + E\left[\|P(t)\|^2 b_\varphi^2\|\bar{d}(\tau)\|^2\right]\right)\\ &= \sum_{\tau=-\infty}^{t}\mu^{2(t-\tau)}E\left[\|P(t)\|^2\right] \times\left([b_\varphi^4/(1-\mu)^2]\Delta^2 + b_\varphi^2\sigma^2\right).\end{aligned}$$

Inserting in this expression the bound for $\|P(t)\|_{L^2}$ given in the Proposition, one finally gets

$$\|\tilde{\vartheta}(t)\|_{L^2}^2 \leq c'\lambda + c''\frac{1}{\lambda} \quad (13)$$

where c' and c'' are linearly depend on Δ^2 and σ^2 respectively, and the inequality is valid for sufficiently large values of the algorithm memory length.

The applicability of the results worked out in this section can be suitably evaluated by considering the examples introduced in the previous section. For the linear combiner model, condition (10), and hence A.4, is met whenever the input signal $u_i(\cdot)$, $i = 1, 2, \ldots, n$, have a sufficiently rich distribution probability. It can be readily seen that this happens, for example, if $u_i(\cdot)$ are white Gaussian noise ($u_i(\cdot) \sim \mathrm{WGN}(0, \lambda_i^2)$, $\lambda_i^2 > 0$) independent of each other. Then, the bound (13) holds true if the system variables $\varphi(\cdot)$, $\delta\vartheta^\circ(\cdot)$ and $d(\cdot)$ satisfy A.1–A.3, which are conditions that do not look particularly restrictive in the linear combiner case. On the contrary, the facile framework of this section does not suit the situations described in Example 2 (Autoregressive model). In fact, in this case, the observation vector at time t depends on all the past values of the disturbance and the parameter drift so that Assumption A.3 is not met. It is also instructive to discuss the above assumptions in connection with the Hammerstein model. In particular, this allows one to better understand the role played by Assumption A.4. Suppose, for instance, that

$$u(t) = \begin{cases} u'(t), & \text{if } |u'(t)| \leq k \\ k \text{ sign } (u'(t)), & \text{otherwise}\end{cases}$$

where k is a given constant, and $u'(\cdot) \sim \mathrm{WGN}(0, \lambda^2)$ (note that considering a "cut input" is reasonable in connection with nonlinear gain with saturation effects). Then, by making β orthogonal to vector $[1\ k\ k^2 \cdots k^n]'$, it is easily seen that condition (10) is not met; therefore, Assumption A.4 turns out to be too tight in this case.

The above considerations show that the assumptions introduced in this section are too stiff to be applicable to many situations of interest. In the next section, we will introduce a wider analysis framework and, by some additional work, we will show that the dependence law (13) still holds true. The resulting theory looks quite powerful and can be applied, for instance, to Examples 2 and 3.

IV. The Dependent Case

In this section, the following persistent excitation assumption of conditional type will be assumed:

A.5 $\exists s, k_1 > 0, k_2 > 0$ such that

$$p\left(\lambda_{\min}\left\{\sum_{i=\tau+1}^{\tau+s}\bar{\varphi}(i)\bar{\varphi}(i)'\right\} \geq k_1 \,\middle|\, \sigma(\varphi(i), d(i), \delta\vartheta^\circ(i), i \leq \tau)\right) \geq k_2, \forall\tau.$$

Roughly, condition A.5 requires that whatever the past evolution of the system might have been, with probability k_2, the "amount of information" carried by data over the next s time points is greater than k_1 in any direction of the parameter space.

This assumption is much weaker than A.4 in that it does not prevent that information is missing on events with nonzero probability. To show its generality, consider, for instance, the Hammerstein model introduced in Section I (Example 3), and assume that it is fed by

$$u(t) = \begin{cases} u'(t), & \text{if } |u'(t)| \leq k \\ k \text{ sign } (u'(t)), & \text{otherwise}\end{cases}$$

$k > 0$ given constant, and $u'(\cdot) \sim \mathrm{WGN}(0, \lambda^2)$ and $\lambda^2 > 0$, independent of $\sigma(\delta\vartheta^\circ(\cdot), d(\cdot))$. Then, A.5 reduces to $p(\lambda_{\min}\{\sum_{i=\tau+1}^{\tau+s}\bar{\varphi}(i)\bar{\varphi}(i)'\} \geq k_1) \geq k_2, \forall\tau$. This last condition is trivially verified by observing that $E[\bar{\varphi}(i)\bar{\varphi}(i)'] \geq A > 0$ and that $\bar{\varphi}(\cdot)$ is a sequence of independent random variables. In addition, the case of autoregressive processes (Example 2 in Section I) can be studied under assumption A.5. The corresponding analysis, however, is much more complicated than that for the Hammerstein model and is reported in a forthcoming paper.

Theorem: Under assumptions A.1, A.2, and A.5, there exist $\bar{\lambda}$ such that for any $\lambda \geq \bar{\lambda}$

$$\|\tilde{\vartheta}(t)\|_{L^2}^2 \leq c_1\Delta^2\lambda + c_2\sigma^2\frac{1}{\lambda} \quad (14)$$

where c_1 and c_2 are constant, which depends on s, k_1, and k_2 only.

Proof: For notational convenience, we introduce the following two σ algebras:

$$\mathcal{P}^{t-ms} = \sigma(\varphi(j), j \le t-ms; d(t-ks), \delta\vartheta^\circ(t-ks), k \ge m), \quad m \ge 0,$$
$$\mathcal{F}^{t-ms} = \sigma(\varphi(j), j \le t-ms; d(t-ks), \delta\vartheta^\circ(t-ks), k \ge m+1), \quad m \ge 0.$$

Assumptions A.1 and A.2, respectively, entail

$$d(t-ms) \text{ independent of } \sigma(\mathcal{F}^{t-ms}, \delta\vartheta^\circ(t-ms)), \quad m \ge 0$$
$$\delta\vartheta^\circ(t-ms) \text{ independent of } \sigma(\mathcal{F}^{t-ms}, d(t-ms)), \quad m \ge 0,$$

so that the σ algebra $\mathcal{F}^{t-ms}$ turns out to be independent of $\sigma(d(t-ms), \delta\vartheta^\circ(t-ms))$, $m \ge 0$.

Point 1: Construction of an independent set of events $\{A_j\}_{j\ge 0}$ such that

A) $A_j! \subseteq \left\{\lambda_{\min}\left(\sum_{i=t-(j+1)s+1}^{t-js} \bar\varphi(i)\bar\varphi(i)'\right) \ge k_1\right\}$, $j \ge 0$
B) $p(A_j) = k_2$, $j \ge 0$
C) $\sigma(\mathcal{F}^{t-ms}; \{A_j\}_{j\ge 0})$ independent of $\sigma(d(t-ms), \delta\vartheta^\circ(t-ms))$, $m \ge 0$.

Note first that property C) is met if $\{A_j\}_{0\le j\le n}$ satisfies, for any n, the following two conditions (see the Appendix for the proof of this claim):

i) $\sigma(\mathcal{F}^{t-ms};\ A_j,\ j = m, m+1, \dots n)$ independent of $\sigma(d(t-ms), \delta\vartheta^\circ(t-ms))$, $0 \le m \le n$;
ii) $\sigma(\mathcal{P}^{t-ms};\ A_j,\ j = m, m+1, \dots n)$ independent of $\sigma(A_{m-1})$, $1 \le m \le n+1$.

The construction of the set of events $\{A_j\}_{j\ge 0}$ is performed recursively. Then, suppose that the $\{A_j\}_{0\le j\le n}$ satisfy A) and B) with $j \le n$ and i) and ii). For the definition of $A_{n+1} \subseteq \left\{\lambda_{\min}\left(\sum_{i=t-(n+1)s}^{t-(n+2)s+1} \bar\varphi(i)\bar\varphi(i)'\right) \ge k_1\right\}$, first impose that

$$p(A_{n+1}|\mathcal{P}^{t-(n+2)s}) = k_2 \tag{15}$$

(note that this is possible in view of Assumption A.5, provided that the probability space is sufficiently rich). Equation (15) defines the conditional distribution of event A_{n+1} with respect to $\mathcal{P}^{t-(n+2)s}$ and implies condition B) with $j = n+1$. Next, for $m = n+1, n, n-1, \dots, 1, 0$, recursively impose the following double independence condition:

$$\sigma(A_{n+1}) \text{ independent of } \sigma(d(t-ms), \delta\vartheta^\circ(t-ms)) \text{ conditionally to } \sigma(\mathcal{F}^{t-ms}; A_j, j=m, m+1, \dots, n) \tag{16}$$

and

$$\sigma(A_{n+1}) \text{ independent of } \sigma(A_{m-1}) \text{ conditionally to } \sigma(\mathcal{P}^{t-ms}; A_j, j=m, m+1, \dots, n). \tag{17}$$

For $m = n+1$, condition (16) can be imposed because of the independence between $\mathcal{F}^{t-(n+1)s}$ and $\sigma(d(t-(n+1)s), \delta\vartheta^\circ(t-(n+1)s))$. Condition (16), with $m \le n$ and (17) can be imposed in view of the fact that $\{A_j\}_{0\le j\le n}$ satisfies i) and ii).

From (15)–(17), properties i) and ii) *with* $n+1$ *in place of* n follow (in the Appendix, a detailed proof of this fact is given). This ends point 1.

By means of the sequence $\{A_j\}_{j\ge 0}$, we can now construct the following matrix (fake information matrix):

$$P_1(t)^{-1} = \sum_{j=0}^{\infty} \mu^{(j+1)s-1} k_1 I_{A_j},$$
$$\text{where } I_{A_j} = \begin{cases} \text{identity matrix on } A_j \\ 0 \text{ otherwise.} \end{cases}$$

Note that property C) of $\{A_j\}_{j\ge 0}$ implies that $P_1(t)$ is independent of $\sigma(d(t-ks), \delta\vartheta^\circ(t-ks), k \ge 0)$. Moreover, from property A), it follows that $P_1(t)^{-1} \le P(t)^{-1}$.

Point 2—Bounding $E[\|P_1(t)\|^2]$*:* Denoting by 1_A the indicator function associated with the set A and taking into account that events $\{A_j\}_{j\ge 0}$ constitute an independence sequence, one has

$$\begin{aligned}
&E[\|P_1(t)\|^2] \\
&= E\left[\left(\sum_{j=0}^{\infty} \mu^{(j+1)s-1} k_1 1_{A_j}\right)^{-2}\right] \\
&= E\left[\left(\mu^{s-1} k_1 1_{A_0} + \mu^s \sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right] \\
&= k_2 \mu^{-2s} E\left[\left(\mu^{-1} k_1 + \sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right] \\
&\quad + (1-k_2)\mu^{-2s} E\left[\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right] \\
&= k_2 \mu^{-2s} E\left[\frac{\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}}{\left(1 + \mu^{-1} k_1 \left[\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right]^{1/2}\right)^2}\right] \\
&\quad + (1-k_2)\mu^{-2s} E\left[\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right] \\
&\le k_2 \mu^{-2s} \frac{E\left[\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right]}{\left(1 + \mu^{-1} k_1 \left\{E\left[\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right]\right\}^{1/2}\right)^2} \\
&\quad + (1-k_2)\mu^{-2s} E\left[\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right] \\
&\quad \text{(using jensen's inequality).}
\end{aligned} \tag{18}$$

Observing that

$$E\left[\left(\sum_{j=0}^{\infty} \mu^{(j+1)s-1} k_1 1_{A_j}\right)^{-2}\right] = E\left[\left(\sum_{j=1}^{\infty} \mu^{js-1} k_1 1_{A_j}\right)^{-2}\right]$$

from (18), one immediately has

$$(\mu^{2s}-1+k_2)\mu^{-2s-2}k_1^2E\left[||P_2(t)||^2\right] + 2\mu^{-2s-1}k_1(\mu^{2s}-1+k_2) \left(E\left[||P_1(t)||^2\right]\right)^{1/2}-\mu^{-2s}+1\le 0$$

so that there exists $k<\infty$ such that

$$\limsup_{\mu\to 1} E\left[||P_1(t)||^2\right]/(1-\mu)^2 \le K. \tag{19}$$

Point 3—Bounding $\|\tilde{\vartheta}(t)\|_{L^2}$: The determination of the bound for $\tilde{\vartheta}(\cdot)$ is performed by replacing $P(t)$ with the auxiliary covariance matrix $P_1(t)$. This allows one to resort to standard probabilistic techniques such as those used when $\varphi(\cdot)$ is independent of $\delta\vartheta^\circ(\cdot)$ and $d(\cdot)$.

We start by considering the effect of the disturbances $\{d(t-ks), k\ge 0\}$ and the drift terms $\{\delta\vartheta^\circ(t-ks), k\ge 0\}$ on the tracking error $\tilde{\vartheta}(t)$. The effect of $\{d(t-ks-i), k\ge 0\}$ and $\{\delta\vartheta^\circ(t-ks-i), k\ge 0\}$, $i=1,2,\ldots,s-1$ can be studied in the same way.

From property A) of $\{A_j\}_{j\ge 0}$, it follows that $P_1(t)^2 \ge P(t)^2$. Consequently, the effect of $\{d(t-ks), k\ge 0\}$ and $\{\delta\vartheta^\circ(t-ks), k\ge 0\}$ on the L^2 norm of the parameter estimation error $\tilde{\vartheta}(t)$ can be given an upper bound as follows (see (9)):

$$\begin{aligned}
&E\left[\left\|\sum_{k=0}^{\infty}\mu^{ks}P(t)\left[-P(t-ks)^{-1}\delta\vartheta^\circ(t-ks)\right.\right.\right.\\
&\qquad\left.\left.\left.+\ \bar{\varphi}(t-ks)\bar{d}(t-ks)\right]\right\|^2\right]\\
&= E\left[\left\{\sum_{k_1=0}^{\infty}\mu^{k_1s}\left[-\delta\vartheta^\circ(t-k_1s)'P(t-k_1s)^{-1}\right.\right.\right.\\
&\qquad\left.\left.\left.+\ \bar{d}(t-k_1s)\bar{\varphi}(t-k_1s)'\right]\right\}P(t)^2\right.\\
&\qquad\left.\times\left\{\sum_{k_2=0}^{\infty}\mu^{k_2s}\left[-P(t-k_2s)^{-1}\delta\vartheta^\circ(t-k_2s)\right.\right.\right.\\
&\qquad\left.\left.\left.+\ \bar{\varphi}(t-k_2s)\bar{d}(t-k_2s)\right]\right\}\right]\\
&\le E\left[\left\{\sum_{k_1=0}^{\infty}\mu^{k_1s}\left[-\delta\vartheta^\circ(t-k_1s)'P(t-k_1s)^{-1}\right.\right.\right.\\
&\qquad\left.\left.\left.+\ \bar{d}(t-k_1s)\bar{\varphi}(t-k_1s)'\right]\right\}P_1(t)^2\right.\\
&\qquad\left.\times\left\{\sum_{k_2=0}^{\infty}\mu^{k_2s}\left[-P(t-k_2s)^{-1}\delta\vartheta^\circ(t-k_2s)\right.\right.\right.\\
&\qquad\left.\left.\left.+\ \bar{\varphi}(t-k_2s)\bar{d}(t-k_2s)\right]\right\}\right].
\end{aligned}\tag{20}$$

By the independence property C), it turns out that the only terms to be nonzero in the right-hand side of (20) are the synchronous terms containing either $\delta\vartheta^\circ(\cdot)$ or $\bar{d}(\cdot)$. Let us show, for instance, that the term that contains $\delta\vartheta^\circ(t-k_1s)$ and $\delta\vartheta^\circ(t-k_2s)(k_1<k_2)$ equals zero:

$$\begin{aligned}
&E\left[\delta\vartheta^\circ(t-k_1s)'P(t-k_1s)^{-1}P_1(t)^2\right.\\
&\quad\left.\times P(t-k_2s)^{-1}\delta\vartheta^\circ(t-k_2s)\right]\\
&= E\left[E[\delta\vartheta^\circ(t-k_1s)'P(t-k_1s)^{-1}P_1(t)^2P(t-k_2s)^{-1}\right.\\
&\quad\left.\times\delta\vartheta^\circ(t-k_2s)|\sigma(\mathcal{F}^{t-k_1s};\{A_j\}_{j\ge 0})]\right]\\
&= E\left[E[\delta\vartheta^\circ(t-k_1s)'|\sigma(\mathcal{F}^{t-k_1s};\{A_j\}_{j\ge 0})]P(t-k_1s)^{-1}\right.\\
&\quad\left.\times P_1(t)^2P(t-k_2s)^{-1}\delta\vartheta^\circ(t-k_2s)\right]\\
&= E\left[E[\delta\vartheta^\circ(t-k_1s)']P(t-k_1s)^{-1}P_1(t)^2P(t-k_2s)^{-1}\right.\\
&\quad\left.\times\delta\vartheta^\circ(t-k_2s)\right]\\
&= 0.
\end{aligned}$$

Now, observe that property C) entails that $P_1(t)$ is independent of $\sigma(d(t-ks),\delta\vartheta^\circ(t-ks), k\ge 0)$. Thus, in fully analogy with the independent case, from inequality (20), one gets

$$\begin{aligned}
&E\left[\left\|\sum_{k=0}^{\infty}\mu^{ks}P(t)\left[-P(t-ks)^{-1}\delta\vartheta^\circ(t-ks)\right.\right.\right.\\
&\qquad\left.\left.\left.+\ \bar{\varphi}(t-ks)\bar{d}(t-ks)\right]\right\|^2\right]\\
&\le\sum_{k=0}^{\infty}\mu^{2ks}E\left[\|P_1(t)\|^2\right]\left([b_\varphi^4/(1-\mu)^2]\Delta^2+b_\varphi^2\sigma^2\right)
\end{aligned}\tag{21}$$

Inserting in (21) the bound for $E[\|P_1(t)\|^2]$ given by (19), one can finally conclude that for μ sufficiently close to 1, the effect of $\{d(t-ks),\delta\vartheta^\circ(t-ks), k\ge 0\}$ on the L^2 norm of $\tilde{\vartheta}(t)$ is bounded by the quantity

$$2Kb_\varphi^4\Delta^2/(1-\mu^{2s})+2Kb_\varphi^2\sigma^2(1-\mu)^2/(1-\mu^{2s}).$$

Handling in an analogous way $\{d(t-ks-i),\delta\vartheta^\circ(t-ks-i), k\ge 0\}$, $i=1,2,\ldots,s-1$, the statement of the theorem follows. □

Remark: Note that according to (14), if $\Delta^2=0$ (no drift in the true parameterization), the L^2 norm of the estimation error tends to zero as the forgetting factor tends to 1. In other words, the parameter estimate approaches the true parameterization at will, provided that the memory length of the algorithm is taken sufficiently long. □

V. Conclusion

One of the crucial points in the analysis of identification algorithms is the description of how stochastic information can compensate for uncertainty in the parameter estimation. Reportedly, this task turns out to be particularly hard for RLS-type algorithms, especially in the truly dynamic case, when the observation vector depends on the past history of both disturbance and parameter drift. In this paper, a new approach is proposed to cope with this problem. The basic idea consists of "cleaning" the information pattern by its "dependent components." Then, the analysis is carried out by focusing on the "independent information component" of data. It is interesting to note that under mild excitation assumptions of conditional type, the independent component turns out to be rich enough to guarantee good tracking performance.

$$\left\{E\left[(\lambda_{\min}\{\Phi(\tau-s+1,\tau)\}+\mu^s\lambda_{\min}\{\Phi(-\infty,\tau-s)\})^{-2}\right]\right\}^{-1/2}$$

$$=\left\{E\left[\frac{(\lambda_{\min}\{\Phi(\tau-s+1,\tau)\})^{-2}(\mu^s\lambda_{\min}\{\Phi(-\infty,\tau-s)\})^{-2}}{\left(\left\{(\lambda_{\min}\{\Phi(\tau-s+1,\tau)\})^{-2}\right\}^{1/2}+\left\{(\mu^s\lambda_{\min}\{\Phi(-\infty,\tau-s)\})^{-2}\right\}^{1/2}\right)^2}\right]\right\}^{-1/2}$$

$$\geq\left\{\frac{E\left[(\lambda_{\min}\{\Phi(\tau-s+1,\tau)\})^{-2}\right]E\left[(\mu^s\lambda_{\min}\{\Phi(-\infty,\tau-s)\})^{-2}\right]}{\left(\left\{E\left[(\lambda_{\min}\{\Phi(\tau-s+1,\tau)\})^{-2}\right]\right\}^{1/2}+\left\{E\left[(\mu^s\lambda_{\min}\{\Phi(-\infty,\tau-s)\})^{-2}\right]\right\}^{1/2}\right)^2}\right\}^{-1/2}.$$

The proposed approach allows one to extend to RLS-type algorithms for the identification of systems with dependent regressors the following fundamental law, which is well known for LMS-type algorithms:

$$\|\hat{\vartheta}(t)-\vartheta^\circ(t)\|_{L^2}^2\leq c'\lambda+c''\frac{1}{\lambda},\quad \lambda=\text{memory length}.$$

We end the paper by indicating two directions worthy of further research.

i) Equation (14) gives an upper bound for the estimation error that explicitly depends on the memory length of the algorithm and the variance of the parameter drift and the noise. Moreover, based on the proof of the theorem, it is not hard to work out explicit expressions for the constants c_1 and c_2 as functions of s, k_1, and k_2. On the contrary, more work is required in order to state explicitly the dependence of the error on the size of the observation vectors. This is a main issue for applications where the selection of the dimension of regressors plays often a crucial rule.

ii) In this paper, the theory has been developed for exponentially weighted LS identification algorithms only. Extensions to more general classes of LS algorithms are expected.

APPENDIX

PROOF OF (11):

Using inequality $\lambda_{\min}\{A+B\}\geq\lambda_{\min}\{A\}+\lambda_{\min}\{B\}$, $A\geq 0$, $B\geq 0$, the left-hand side of (11) can be given a lower bound as follows:

$$\left\{E\left[(\lambda_{\min}\{\Phi(-\infty,\tau)\})^{-2}\right]\right\}^{-1/2}$$
$$\geq\left\{E\left[(\lambda_{\min}\{\Phi(\tau-s+1,\tau)\}+\mu^s\lambda_{\min}\{\Phi(-\infty,\tau-s)\})^{-2}\right]\right\}^{-1/2}. \quad (22)$$

Applying twice the Jensen's inequality to the right-hand side of (22) as well as taking into account the independence between $\Phi(\tau-s+1,\tau)$ and $\Phi(-\infty,\tau-s)$, one obtains the expressions at the top of the page. This last expression is easily seen to be coincident with the right-hand side of (11). □

PROOF OF THE FACT THAT i), ii) ⇒ C)

From ii), it follows that $\sigma(\mathcal{P}^{t-ms};A_j,\ j=m,m+1,\ldots,n)$ is independent of $\sigma(A_j,j=0,1,\ldots,m-1)$, $1\leq m\leq n$. Hence, also using i), one immediately has that $\sigma(\mathcal{F}^{t-ms};A_j,\ j=0,1,\ldots,n)$ is independent of $\sigma(d(t-ms),\delta\vartheta^\circ(t-ms))$, $0\leq m\leq n$. Observing that $U_n\,\sigma(\mathcal{F}^{t-ms};A_j,j=0,1,\ldots,n)$ is a π algebra that generates $\sigma(\mathcal{F}^{t-ms};\{A_j\}_{j\geq 0})$, property C) straightforwardly follows (see, e.g., vol. 1 of [20] for the concept of π algebra and related results). □

PROOF OF THE FACT THAT (16) ⇒ i) WITH $n+1$ IN PLACE OF n AND (17) + (15) ⇒ ii) WITH $n+1$ IN PLACE OF n

We only show that (16) ⇒ i). The implication (17) + (15) ⇒ ii) can be shown in a similar way.

For a given $m\in[0,n+1]$, consider the set of events

$$\mathcal{S}=\{A_{n+1}\cap F,\bar{A}_{n+1}\cap F;\ F\in\sigma(\mathcal{F}^{t-ms};A_j,j=m,m+1,\ldots n)\}.$$

$\mathcal{S}$ is independent of $\sigma(d(t-ms),\delta\vartheta^\circ(t-ms))$. Indeed, for any $B\in\sigma(d(t-ms),\delta\vartheta^\circ(t-ms))$, one has

$$\begin{aligned}
&p((A_{n+1}\cap F)\cap B)\\
&=E\Big[p\Big((A_{n+1}\cap F)\cap B\Big|\sigma(\mathcal{F}^{t-ms};A_j,j=m,m+1,\\
&\qquad\ldots n;d(t-ms),\delta\vartheta^\circ(t-ms))\Big)\Big]\\
&=E\Big[p\Big(A_{n+1}\Big|\sigma(\mathcal{F}^{t-ms};A_j,j=m,m+1,\ldots n;\\
&\qquad d(t-ms),\delta\vartheta^\circ(t-ms))\Big)I_{F\cap B}\Big]\\
&=E\Big[p\Big(A_{n+1}\Big|\sigma(\mathcal{F}^{t-ms};A_j,j=m,m+1,\ldots n)\Big)I_{F\cap B}\Big]\\
&\text{(using (16))}\\
&=E\Big[p\Big(A_{n+1}\cap F\Big|\sigma(\mathcal{F}^{t-ms};A_j,j=m,m+1,\ldots n)\Big)I_B\Big]\\
&=E\Big[p\Big(A_{n+1}\cap F\Big|\sigma(\mathcal{F}^{t-ms};A_j,j=m,m+1,\ldots n)\Big)\Big]\\
&\quad\times p(B)\\
&=p(A_{n+1}\cap F)p(B),
\end{aligned}$$

and, analogously $p((\bar{A}_{n+1} \cap F) \cap B) = p(\bar{A}_{n+1} \cap F)p(B)$. Observing that $\mathcal{S}$ is a π algebra that generates $\sigma(\mathcal{F}^{t-ms}; A_j, j = m, m+1, \ldots, n+1)$, the thesis follows.

REFERENCES

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

[2] G. E. P. Box and G. M. Jenkins, *Time Series Analysis*. San Francisco: Holden Day, 1970.

[3] K. S. Narendra and P. G. Gallman, "An iterative method for the identification of nonlinear systems using a Hammerstein model," *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 546–550, 1966.

[4] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.

[5] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*. Englewood Cliffs, NJ: Prentice Hall, 1984.

[6] T. R. Fortescue, L. S. Kershenbaum, and B. E. Ydstie, "Implementation of self-tuning regulators with variable forgetting factors," *Automatica*, vol. 17, pp. 831–835, 1981.

[7] D. Bertin, S. Bittanti, and P. Bolzern, "Tracking of nonstationary systems by means of different prediction error directional forgetting techniques," in *Proc. 2nd IFAC Workshop Adaptive Systems Contr. Signal Processing* (Lund, Sweden), 1986, pp. 91–96.

[8] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[9] D. T. Lee, M. Morf, and B. Friedlander, "Recursive least squares ladders estimation algorithms," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-29, pp. 627–641, 1981.

[10] O. Macchi, "Optimization of adaptive identification for time-varying filters," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 283–287, 1986.

[11] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, "Stationary and nonstationary learning characteristics of the LMS adaptive filters," *Proc. IEEE*, vol. 64, no. 8, pp. 1151–1162, 1976.

[12] E. Eleftheriou and D. D. Falconer, "Tracking properties and steady-state performance of RLS adaptive filter algorithms," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-34, pp. 1097–1109, 1986.

[13] T. Adali and S. H. Ardalan, "Fixed-point roundoff error analysis of the RLS algorithm with time-varying channels," *IEEE Int. Conf. Acoustic, Speech, Signal Processing* (Toronto), 1991, pp. 1865–1868.

[14] R. L. Lozano, "Identification of time-varying linear models," in *Proc. 22nd Conf. Decision Contr.*, 1983, pp. 604–606.

[15] R. M. Canetti and M. D. España, "Convergence analysis of the least squares identification algorithm with variable forgetting factor for time-varying linear systems," *Automatica*, vol. 25, no. 4, pp. 609–612, 1989.

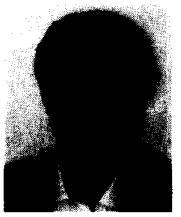
[16] M. Niedzwiecki and L. Guo, "Nonasymptotic results for finite-memory WLS filters," *IEEE Trans. Automat. Contr.*, vol. 36, no. 2, pp. 198–206, 1991.

[17] M. Campi, "Performance of RLS identification algorithms with forgetting factor—A Φ-mixing approach," to appear in *J. Math. Syst., Estimation Contr.*

[18] S. Bittanti and M. Campi, "Adaptive RLS algorithms under stochastic excitation — L^2-convergence analysis," *IEEE Trans. Automat. Contr.*, vol. 36, no. 8, pp. 963–967, 1991.

[19] ——, "Adaptive RLS algorithms under stochastic excitation—Strong consistency analysis," *Syst. Contr. Lett.*, no. 17, pp. 3–8, 1991.

[20] I. I. Gihman and A. V. Skorohod, *The Theory of Stochastic Processes*. Englewood Cliffs, NJ: Prentice Hall, 1974.



Marco C. Campi was born in Varese, Italy, on December 7, 1963. He received the Doctor degree (Laurea) in electrical engineering (summa cum laude) from the Politecnico di Milano in 1988.

From 1988 to 1992, he worked as a researcher at the Centro di Studio per la Teoria del Sistemi of the Consiglio Nazionale della Ricerche, and during 1992, he was a visiting researcher at the Australian National University in Canberra. Currently, he is an associate professor in the Department of Electrical Engineering and Automation at the University of Brescia. His main research interests are in the areas of adaptive identification and control, prediction theory, and robust control.