Bounded Error Identification of Time-Varying Parameters by RLS Techniques

Sergio Bittanti and Marco Campi

Abstract—**The performance of the Recursive Least Squares algorithm with constant forgetting factor in the identification of time-varying parameters is studied in a stochastic framework. It is shown that the mean square tracking error keeps bounded if and only if the so-called covariance matrix of the algorithm is L^1-bounded. Then, a feasibility range for the forgetting factor is worked out in correspondence of which the covariance matrix (and therefore the tracking error) keeps bounded.**

I. Introduction

A. The RLS Algorithm for the Tracking of Time-Varying Parameters

The challenge of adaptive identification techniques is to allow for good performance in prediction, filtering, and control despite possible changes in the system dynamics.

In this context, a typical setting of analysis amounts to assuming that the system is described by the equations

$$y(t) = \vartheta^0(t)'\varphi(t) + d(t) \tag{1.1a}$$

$$\vartheta^0(t+1) = \vartheta^0(t) + \delta\vartheta^0(t) \tag{1.1b}$$

where $d(t)$ is the system disturbance and $\delta\vartheta^0(t)$ the parameter drift. The time-varying parameter $\vartheta^0(t) \in \mathbf{R}^{n\times m}$ has to be estimated starting from the measurements of the output $y(t) \in \mathbf{R}^m$ and the observation vector $\varphi(t) \in \mathbf{R}^n$ up to time t. To this purpose, a major role is played by the Recursive Least Squares (RLS) algorithm, wherein adaptivity is often achieved by means of the so-called forgetting factor (FF) (see [1] and [2]). More precisely, the estimate $\hat{\vartheta}(t)$ of the unknown parameter is obtained by the equations

$$\epsilon(t) = v(t)' - \psi(t)'\hat{\vartheta}(t-1) \tag{1.2a}$$

$$P(t) = \frac{1}{\mu}\left[P(t-1) - \frac{P(t-1)\psi(t)\psi(t)'P(t-1)}{1+\psi(t)'P(t-1)\psi(t)}\right] \tag{1.2b}$$

$$K(t) = \mu P(t)\psi(t) \tag{1.2c}$$

$$\hat{\vartheta}(t) = \hat{\vartheta}(t-1) + K(t)\epsilon(t). \tag{1.2d}$$

In these expressions, vectors $\psi(t) \in \mathbf{R}^n$, $v(t) \in \mathbf{R}^m$ are filtered versions of the measurements $\varphi(t)$ and $y(t)$, respectively. ((1.2b) and (1.2d) are initialized at time $t = 0$ with deterministic matrices $P(0) = P(0)' > 0$ and $\hat{\vartheta}(0)$, respectively.) Note that (1.2b) can be equivalently written as

$$P(t)^{-1} = \mu P(t-1)^{-1} + \mu\psi(t)\psi(t)'. \tag{1.3}$$

The specific choice of μ has been discussed in many papers, such as [3] and [4].

If $\psi(t)$ is let to coincide with the observation vector $\varphi(t)$ and $v(t)$ with the output vector $y(t)$, then one obtains the Exponential Forgetting (EF) algorithm, which has been the subject of a number of papers, see e.g., [3] and [5]. However, it is well known that EF may exhibit poor performance when $\varphi(\cdot)$ takes large values. Indeed, in such a case, $P(t)^{-1}$ becomes exceedingly large in the $\varphi(\cdot)$-directions, (see (1.3)), and this results in a reduction of the algorithm sensitivity in these directions. Such a reduced sensitivity has some persistence in time since the recovery of $P(\cdot)$ cannot be prompt (see (1.2b)). In these periods of sluggishness, the estimate becomes very poor if there is a permanent drift of the system parameters in the directions of low sensitivity. To overcome effects of this type, one can feed algorithm (1.2) with vectors $\psi(t)$ and $v(t)$ derived as

$$\psi(t) = \begin{cases} \varphi(t), & \|\varphi(t)\| \le H \\ \frac{H}{\|\varphi(t)\|}\varphi(t), & \|\varphi(t)\| > H \end{cases}$$

$$v(t) = \begin{cases} y(t), & \|\varphi(t)\| \le H \\ \frac{H}{\|\varphi(t)\|}y(t), & \|\varphi(t)\| > H. \end{cases} \tag{1.4}$$

The value H of the threshold is a user's choice. A typical guideline is to take H high enough with respect to the normal ranges spanned by the variables entering vector $\varphi(t)$. In this way, the "cut" of raw data is active only when the observation vector becomes exceedingly large, as may happen due to oversized shots of noise, outliers and so on. Precautions such as (1.4) are frequently encountered in the area of robust estimation under the heading of bounded influence regression, see [6] and [2] for more discussion.

B. From the Deterministic to the Stochastic Analysis

Letting $\tilde{\vartheta}(t) = \hat{\vartheta}(t) - \vartheta^0(t+1)$, from (1.1), (1.2a), (1.2d) one obtains

$$\tilde{\vartheta}(t) = F(t)\tilde{\vartheta}(t-1) + K(t)n(t)' - \delta\vartheta^0(t), \tag{1.5a}$$

$$F(t) = I - K(t)\psi(t)' \tag{1.5b}$$

$$n(t) = \begin{cases} d(t), & \|\varphi(t)\| \le H \\ \frac{H}{\|\varphi(t)\|}d(t), & \|\varphi(t)\| > H. \end{cases}$$

Equation (1.5) describes the influence of the drift term $\delta\vartheta^0(\cdot)$ and the disturbance term $d(\cdot)$ on the parameter estimation error $\tilde{\vartheta}(\cdot)$ and forms the basis of all the performance analyses. As pointed out in [7], the difficulty in the analysis stems from the complicated expression for the system transition matrix

$$\Phi(t, \tau) = F(t)F(t-1)\cdots F(\tau+2)F(\tau+1).$$

Its properties depend entirely on the sequence $\varphi(\cdot)$, but they are inherited in a fairly complicated way. Obviously, the uniform exponential stability of the above time-varying linear system is a basic desirable property. This amounts to requiring

$$\|\Phi(t, \tau)\| \le C\rho^{t-\tau}, \qquad |\rho| < 1. \tag{1.6}$$

Then, if $K(\cdot)$ keeps bounded, the boundedness of $n(\cdot)$ and $\delta\vartheta^0(\cdot)$ entails the boundedness of the estimation error $\tilde{\vartheta}(\cdot)$. Condition (1.6) imposes a deterministic contractivity property. As such, it calls for some deterministic excitation assumption on data. The typical condition takes the form

$$0 < cI \le \sum_{i=\tau+1}^{\tau+s} \varphi(i)\varphi(i)' \le CI, \quad \forall\tau \tag{1.7}$$

Manuscript received October 21, 1991; revised June 18, 1992 and March 25, 1993. This work was supported by MURST project," Model Identification, System Control, and Signal Processing," and by CNR Project, "Methods and Mathematical Models for Industry."

S. Bittanti is with the Dipartimento di Elettronica, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy.

M. Campi is with the Dipartimento di Elettronica per l'Automazione, Universita' di Brescia, via Branze 38, 25123 Brescia, Italy.

IEEE Log Number 9400354.

0018-9286/94$04.00 © 1994 IEEE

which implies both the uniform stability and the boundedness of the algorithm gain.

This line of analysis can be traced back to [8] where condition (1.7) has been linked to the uniform complete observability and controllability of system (1.5). Ever since, this approach has been adopted in a number of papers and books. As a matter of fact, condition (1.7) constitutes a now common paradigm in system identification. On the other hand, it is obvious that (1.7) is just an idealization far from being applicable to uncertain real data.

Turning to the stochastic analysis, it is worth noticing that most papers are concerned with the case of adaptive algorithms with very long memory length. The stochastic behaviour of the algorithm can then be approximated by means of some deterministic average. A well-established technique to pursue this objective relies on weak convergence concepts leading to the so-called ODE approach, [9]–[11]. More recently, under ergodicity assumption on the observation vector, in [12] the proposal is made to replace the time-varying and stochastic algorithm covariance matrix with a constant and deterministic approximant. The applicability of the corresponding results, however, is limited by the stationarity character of the underlying assumptions. Another noteworthy contribution is the one provided in [13]. Precisely, the estimation error variance is evaluated by squaring (1.5.a) and then simplifying the obtained expression by dropping out the cross terms and by replacing the stochastic matrix $F(t)$ with its expected value. In [13], it is shown that this is a fair approximation when the memory length tends to infinity. The main drawback of this approach is that it calls for constraints on matrix $P(\cdot)$ which look to stiff to be applicable with generality.

In conclusion, there is a huge lack of knowledge in the area of stochastic RLS algorithms and, in most papers in the field, the analytical results are obtained by making reference to a limit case, that of infinite memory length. As a matter of fact, even in elementary situations, such as in the forthcoming example, it is hard to predict the performance obtained by the RLS algorithm for the different values of μ.

Example 1: Suppose that all the variables in (1.1) are scalar and $\varphi(t)$ coincides with an exogenous i.i.d. input $u(t)$ with binomial distribution

$$u(t) = \begin{cases} 2, & \text{prob. } 0.5 \\ 0, & \text{prob. } 0.5. \end{cases}$$

The estimate $\hat{\vartheta}(t)$ can be obtained by minimizing

$$J = \sum_{i=1}^{t} \mu^{t-i}[y(i) - u(i)\hat{\vartheta}(t)]^2 \tag{1.8}$$

where $\mu \in (0, 1)$ is the forgetting factor. Obviously, the estimation error depends on the characteristics of the drift term and the disturbance affecting the signal $y(\cdot)$. Assume, for example, that

$\delta\vartheta^0(\cdot)$ i.i.d. sequence with $L^2 -$ norm $= 10^{-4}$, independent of $u(\cdot)$;

$d(t) =$ const. $= 1$, $\forall t$.

In this way the setting of analysis is fully specified. Notwithstanding the simplicity of the described situation, however, no result in the literature known to the authors allow one to say for which value of μ the tracking error keeps bounded. ■

C. Achievements of the Paper

In this paper, we study the tracking performance of algorithm (1.2) in a stochastic environment without any restriction on the memory length of the algorithm. We will resort to a direct approach requiring neither reformulations nor simplification of any sort.

The main achievements of the paper can be outlined as follows.

a) The second-order moment of the parameter tracking error is shown to be bounded if and only if the expected value of the so called "covariance matrix" $P(\cdot)$ is bounded (Theorem 1).

b) Matrix $P(\cdot)$ can be kept bounded provided that the FF does not fall below a given threshold (which depends on the information content of data) (Proposition 2).

Statement a) has the important consequence that the study of the boundedness of the parameter tracking error reduces to the analysis of the behavior of the sole $P(\cdot)$ equation. In turn, b) points out that, to keep $P(\cdot)$ bounded, one should avoid an overdiscount of the information conveyed by past data. Thus, the threshold of Proposition 2 defines a feasibility range for the forgetting factor. By suitably selecting the value of the forgetting factor within such a range, one can minimize the magnitude of the parameter tracking error.

A highlight of the paper is that the analysis will be based on simple L^2-boundedness conditions on the disturbance $d(\cdot)$ and the drift term $\delta\vartheta^0(\cdot)$ and a general excitation assumption of conditional type on the observation vector $\varphi(\cdot)$. In particular, no assumption on the distribution or stationarity of the underlying processes is made. This is why the results can be applied to a variety of specific situations.

II. Stochastic Setting of Analysis

Given a probability space $(\Omega, \mathcal{F}, p)$ consider the stochastic processes $\varphi(t) \in \mathbb{R}^n$, $d(t) \in \mathbb{R}^m$ and $\delta\vartheta^0(t) \in \mathbb{R}^{n\times m}$. Let $\mathcal{P}_1^t$ be the σ-algebra generated by $(\varphi(i), d(i), \delta\vartheta^0(i) \mid i = 1, 2, \cdots, t)$. As already pointed out in Section I, we will suppose that $y(t)$ and $\vartheta^0(t)$ are recursively generated in agreement with (1.1). Equation (1.1b) is initialized at time $t = 1$ with deterministic initial condition $\vartheta^0(1)$. Then $\mathcal{P}_1^t$ can be seen as the σ-algebra of the past. We also introduce the symbols $\mathcal{P}_1^\infty := \sigma(\bigcup_t \mathcal{P}_1^t)$ and $\mathcal{P}_1^0$ to denote the trivial σ-algebra.

The disturbance $d(t)$ and the drift term $\delta\vartheta^0(t)$ are assumed to be L^2-conditionally bounded:

Assumption A.1: $E[\|d(t+1)\|^2 \mid \mathcal{P}_1^t] \le \Lambda_d^2$, $\forall t \ge 0$, where Λ_d is a deterministic constant. ■

Assumption A.2: $E[\|\delta\vartheta^0(t+1)\|^2 \mid \mathcal{P}_1^t] \le \Lambda_\vartheta^2$, $\forall t \ge 0$, where Λ_ϑ is a deterministic constant. ■

Notice that Assumption A.1 does not require that the noise $d(t)$ have zero expected value. As for the time behaviour of the drift term $\delta\vartheta^0(t)$, Assumption A.2 allows for any deterministic pattern of $\vartheta^0(t)$ with uniformly bounded variations. It allows also for many types of stochastic perturbations provided that the correspondent drift is slow. Thus, Assumption A.2 does not prevent $\vartheta^0(\cdot)$ from the possibility of presenting trends or seasonalities.

In this paper, the excitation condition firstly introduced in [14] is used.

Assumption A.3: There exist $s > 0$ such that

$$\mathfrak{p}\left(\lambda_{\min}\left\{\sum_{i=rs+1}^{(r+1)s} \frac{\varphi(i)\varphi(i)'}{1 + H^{-2}\|\varphi(i)\|^2}\right\} \ge K_1 \,\Big|\, \mathcal{P}_1^{rs}\right) \ge K_2, \quad \forall r \ge 0 \tag{2.1}$$

for some real $K_1 > 0$ and $K_2 > 0$. ■

Although condition (2.1) is equivalent to (2.1) of [15], it is better suited for the technical developments to follow. It 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. In this sense, integer s can be interpreted as excitation horizon.

III. A Necessary and Sufficient Condition for the Boundedness of the Tracking Error

The time-evolution of the tracking error $\tilde{\vartheta}(t) = \hat{\vartheta}(t) - \vartheta^0(t+1)$ is described by (1.5a). To derive a suitable expression for the solution of this equation, observe first that, using (1.5b), (1.2c), and (1.2b)

$$F(t)F(t-1)\cdots F(\tau+2)F(\tau+1) = \mu^{t-\tau}P(t)P(\tau)^{-1}.$$

By means of this expression, the solution of (1.5a) can be given the form

$$\tilde{\vartheta}(t) = \mu^t P(t)P(0)^{-1}\tilde{\vartheta}(0) + \sum_{\tau=1}^{t}\mu^{t-\tau}P(t)P(\tau)^{-1}K(\tau)n(\tau)' - \sum_{\tau=1}^{t}\mu^{t-\tau}P(t)P(\tau)^{-1}\delta\vartheta^0(\tau). \quad (3.1)$$

Equation (3.1) points out the key role played by $E[\mu^{t-\tau}\|P(t)\| \mid \mathcal{P}_1^\tau]$ in determining the error $\tilde{\vartheta}(\cdot)$. Proposition 1 below gives a contractivity result concerning the time behaviour of such a quantity.

Proposition 1: Let $\nu \in [\mu, 1)$ such that $\nu^s > 1 - K_2$. Then

$$E[\mu^{t-\tau}\|P(t)\| \mid \mathcal{P}_1^\tau] \le \nu^{t-\tau}\{(\gamma_{\tau,t}\nu^{2(1-s)}\|P(\tau)\| + (1-\gamma_{\tau,t})\overline{P}(\nu)\}, \qquad \forall t \ge \tau \ge 0$$

where

$$\overline{P}(\nu) = \nu^{1-s}(1-\nu^s)K_1^{-1}(\nu^s - 1 + K_2)^{-1}$$

and $\gamma_{\tau,t} \in (0, 1)$, $\forall\tau, t$, is a deterministic function of τ and t.

Proof: See the Appendix. ■

Consider now any class $\mathcal{C}$ of systems satisfying Assumptions A.1, A.2, A.3, with given constants $\Lambda_d > 0$, $\Lambda_\vartheta > 0$, $s > 0$, $K_1 > 0$, $K_2 > 0$. The theorem below states that the L^2-boundedness of the tracking error $\tilde{\vartheta}(\cdot)$ over such a class is equivalent to the L^1-boundedness of matrix $P(\cdot)$ over the same class.

Theorem 1:

$\mathrm{SUP}_{\mathcal{C}}\mathrm{SUP}_t\|\tilde{\vartheta}(t)\|_{L^2} < \infty \Leftrightarrow \mathrm{SUP}_{\mathcal{C}}\,\mathrm{SUP}_t\|P(t)\|_{L^1} < \infty.$

Proof: $\Rightarrow$) Note first that the forcing term $K(t)n(t)'$ in (1.5a) coincides with $K(t)d(t)'$ if $\|\varphi(t)\| \le H$. Moreover, the disturbance term $d(t)$, being subject to Assumption A.1 only, can be any deterministic vector the norm of which is less than Λ_d. Consequently, it suffices to show that $\sup_{\mathcal{C}}\sup_t\|P(t)\|_{L^1} = \infty \Rightarrow \sup_{\mathcal{C}}\sup_t \int_{\{\|\varphi(t)\|\le H\}}^{\|K(t)\|^2 d\mathfrak{p}} = \infty$.

Given an arbitrary number M, consider a system in $\mathcal{C}$ defined by the triple $\{(\varphi_1(\cdot), d_1(\cdot), \delta\varphi_1^0(\cdot)\}$ such that

$$\|P(\bar{t})\|_{L^1} \ge \mu^{1-s}(4M(1-K_2)^{-1} + H^{-2})$$

for some $\bar{t}$ (note that $K_1 < 1$, otherwise $\sup_{\mathcal{C}}\sup_t\|P(t)\|_{L^1} < \infty$). Let q be the integer part of $\bar{t}/s$.

Consider now a set of deterministic vectors $\{v_i \in \mathbf{R}^n, i = 1, 2, \cdots, s\}$, such that $\lambda_{\min}(\sum_{i=1}^s v_i v_i'/(1 + H^{-2}\|v_i\|^2)) \ge K_1$. Moreover, let $\mathcal{A}$ be an event, independent of the σ-algebra $\mathcal{P}_1^\infty$ associated to the considered system, such that $\mathfrak{p}(\mathcal{A}) = 1 - K_2$ (we assume that such an event exists in our probabilistic space).

By means of $\varphi_1(\cdot)$, $\{v_i\}$ and $\mathcal{A}$, a new sequence of observation vectors $\varphi_2(\cdot)$ can be constructed in the following way:

$$\varphi_2(t) = \varphi_1(t), \qquad t \le qs$$

$$\varphi_2(qs+1) = \begin{cases} \frac{x_{\max}[P(qs)]}{\|P(qs)\|^{1/2}}, & \text{on } \mathcal{A} \\ v_1, & \text{on } \overline{\mathcal{A}} \end{cases}$$

$$\varphi_2(t) = v_{t \bmod s}, \qquad t \ge qs+2$$

where $x_{\max}[\cdot]$ denotes maximum eigenvector.

It is easy to prove that the system defined by the triple $\{\varphi_2(\cdot), d_1(\cdot), \delta\vartheta_1^0(\cdot)\}$ belongs to $\mathcal{C}$ too. Moreover, for such a system,

$$\begin{aligned}
&\int_{\{\|\varphi_2(qs+1)\|\le H\}} \|K(qs+1)\|^2 d\mathfrak{p} \\
&\ge \int_{A\cap\{\|\varphi_2(qs+1)\|\le H\}} \frac{\varphi_2(qs+1)'P(qs)P(qs)\varphi_2(qs+1)}{(1+\varphi_2(qs+1)'P(qs)\varphi_2(qs+1))^2} d\mathfrak{p} \\
&\ge \int_{A\cap\{\|\varphi_2(qs+1)\|\le H\}} \frac{1}{4}\|P(qs)\| d\mathfrak{p} \\
&\ge \int_{A\cap\{\|P(qs)\|\ge H^{-2}\}} \frac{1}{4}\|P(qs)\| d\mathfrak{p} \\
&\ge \frac{1}{4}(1-K_2)(4M(1-K_2)^{-1} + H^{-2} - H^{-2}) = M.
\end{aligned}$$

$\Leftarrow$) Letting

$$v_i(\tau) = P(\tau)^{-1}K(\tau)n(\tau)', \qquad v_2(\tau) = P(\tau)^{-1}\delta\vartheta^0(\tau)$$

for $i = 1, 2$, one obtains ($t \ge \tau$)

$$\begin{aligned}
&E[\|\mu^{t-\tau}P(t)v_i(\tau)\|^2] \\
&\quad\le E[\mu^{2(t-\tau)}\|P(t)\|\,\|v_i(\tau)'P(t)v_i(\tau)\|] \\
&\quad\le E[\mu^{t-\tau}\|P(t)\|\,\|v_i(\tau)'P(\tau)v_i(\tau)\|] \\
&\quad= E[E[\mu^{t-\tau}\|P(t)\| \mid \mathcal{P}_1^\tau]\|v_i(\tau)'P(\tau)v_i(\tau)\|].
\end{aligned}$$

Choose now $\nu \in [\mu, 1)$ such that $\nu^s > 1 - K_2$. Then, by resorting to Proposition 1 and inequality $P(\tau) \le P(\tau-1)/\mu$

$$E[\|\mu^{t-\tau}P(t)v_i(\tau)\|^2] \le E\left[\nu^{t-\tau}\left(\gamma_{\tau,t}\nu^{2(1-s)}\frac{\|P(\tau-1)\|}{\mu} + (1-\gamma_{\tau,t})\overline{P}(\nu)\right)\|v_i(\tau)'P(\tau)v_i(\tau)\|\right]. \quad (3.2)$$

On the other hand, in view of (1.2b), (1.2c), and (1.3), and recalling that $\|\psi(t)\| \le H$, $\forall t$

$$\|n(\tau)K(\tau)'P(\tau)^{-1}K(\tau)n(\tau)'\| \le \mu\|n(\tau)\|^2 \quad (3.3)$$

and

$$\|\delta\vartheta^0(\tau)'P(\tau)^{-1}\delta\vartheta^0(\tau)\| \le \left(\frac{\mu H^2}{1-\mu} + \mu^\tau\|P(0)^{-1}\|\right)\|\delta\vartheta^0(\tau)\|^2. \quad (3.4)$$

Inequalities (3.3) and (3.4) can be used in (3.2) with $i = 1$ and $i = 2$, respectively. The inequalities obtained in this way can be used to find upper bounds for the L^2-norm of the second and third term at the right-hand side of (3.1). As for the first term, by resorting again to Proposition 1, one can easily work out the inequality

$$E[\|\mu^{t-\tau}P(t)P(0)^{-1}\tilde{\vartheta}(0)\|^2] \le \nu^t\max\{\nu^{2(1-s)}\|P(0)\|, \overline{P}(\nu)\}\|\tilde{\vartheta}(0)'P(0)^{-1}\tilde{\vartheta}(0)\|. \quad (3.5)$$

Therefore, thanks to (3.3)–(3.5), and recalling Assumptions A.1, A.2, (3.2) leads to

$$\begin{aligned}
&\|\tilde{\vartheta}(t)\|_{L^2} \\
&\le \nu^{t/2}\max\{\nu^{2(1-s)}\|P(0)\|, \overline{P}(\nu)\}^{1/2}\|\tilde{\vartheta}(0)'P(0)^{-1}\tilde{\vartheta}(0)\|^{1/2} \\
&\quad + \sum_{\tau=1}^{t}\nu^{(t-\tau)/2}\Lambda_d \\
&\quad \cdot E^{1/2}[\gamma_{\tau,t}\nu^{2(1-s)}\|P(\tau-1)\| + (1-\gamma_{\tau,t})\overline{P}(\nu)] \\
&\quad + \sum_{\tau=1}^{t}\nu^{(t-\tau)/2}\Lambda_\varphi \\
&\quad \cdot E^{1/2}\left[\left(\gamma_{\tau,t}\nu^{2(1-s)}\frac{\|P(\tau-1)\|}{\mu} + (1-\gamma_{\tau,t})\overline{P}(\nu)\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&\cdot\left(\frac{\mu H^2}{1-\mu}+\mu^{\tau}\|P(0)^{-1}\|\right)\Bigg]\\
&\le \nu^{t/2}\max\{\nu^{2(1-s)}\|P(0)\|,\overline{P}(\nu)\}^{1/2}\\
&\quad\cdot\|\tilde\vartheta(0)'P(0)^{-1}\tilde\vartheta(0)\|^{1/2}\\
&\quad+\sum_{\tau=1}^{t}\nu^{(t-\tau)/2}\Lambda_d\max\{\nu^{2(1-s)}\|P(\tau-1)\|_{L^1},\,\overline{P}(\nu)\}^{1/2}\\
&\quad+\sum_{\tau=1}^{t}\nu^{(t-\tau)/2}\Lambda_\vartheta\left(\frac{\mu H^2}{1-\mu}+\mu^{\tau}\|P(0)^{-1}\|\right)^{1/2}\\
&\quad\cdot\max\left\{\nu^{2(1-s)}\frac{\|P(\tau-1)\|_{L^1}}{\mu},\,\overline{P}(\nu)\right\}^{1/2}.
\end{aligned}
\tag{3.6}
$$

Since $P(\cdot)$ is L^1-bounded, this inequality immediately leads to the conclusion that $\tilde\vartheta(\cdot)$ is L^2-bounded. ■

Remark: Note that, in the noise-free case ($d(\cdot)=0$), to guarantee the boundedness of the estimation error, the boundedness of matrix $P(\cdot)$ is no more necessary. In [16], one can find a simple deterministic example where $\tilde\vartheta(\cdot)$ keeps bounded even if $P(\cdot)$ tends to infinity. ■

It is obvious that matrix $P(\cdot)$ may keep bounded or not depending on the value of the memory length of the algorithm. Indeed, the information conveyed by fresh data may not be sufficient to compensate for the discount of past information if the forgetting factor is exceedingly small. A feasible range for the forgetting factor such that matrix $P(\cdot)$ keeps bounded is given in the following

Proposition 2: Suppose that $\mu^s > 1 - K_2$. Then

$$\underset{C}{\text{SUP}}\,\underset{t}{\text{SUP}}\|P(t)\|_{L^1}\le\max\{\mu^{1-s}\|P(0)\|,\quad\overline{P}(\mu)\}$$

where

$$\overline{P}(\mu)=\mu^{1-s}(1-\mu^s)K_1^{-1}(\mu^s-1+K_2)^{-1}.$$

Proof: See the Appendix. ■

Theorem 1 and Proposition 2 immediately lead to a feasible range for the forgetting factor.

Theorem 2: $\mu^s > 1-K_2 \Rightarrow \text{SUP}_C\,\text{SUP}_t\|\tilde\vartheta(t)\|_{L^2}<\infty$. ■

Example 1 (continued): Consider algorithm (1.2), (1.4) and take $H=2$. With this choice $\psi(t)=\varphi(t)=u(t)$, and $v(t)=u(t)$, $\forall t$, so that algorithm (1.2), (1.4) reduces to the usual EF algorithm. On the other hand, it is well known that EF provides a recursive way to minimize loss function (1.8). Therefore, the results derived in this paper are applicable to the situation described in Example 1.

Sequences $\delta\vartheta^0(\cdot)$ and $d(\cdot)$ obviously satisfy Assumptions A.1 and A.2 with $\Lambda_\vartheta = 10^{-2}$ and $\Lambda_d = 1$, respectively. Moreover, the observation sequence $\varphi(\cdot)=u(\cdot)$ satisfies the excitation condition stated in Assumption A.3 with $s=1$, $K_1=2$, $K_2=1/2$. Indeed

$$1/2=\mathfrak{p}\left(\frac{u(i)^2}{1+u(i)^2/4}\ge 2\right)=\mathfrak{p}\left(\frac{u(i)^2}{1+u(i)^2/4}\ge 2\,\middle|\,\mathcal{P}_1^{i-1}\right).$$

Then, Theorem 2 entails that the tracking error $\tilde\vartheta(\cdot)$ keeps bounded whenever $\mu > 1/2$. ■

Appendix
(Proofs of Proposition 1 and Proposition 2)

The following lemma (the proof of which is omitted) will be used in the subsequent derivations.

Lemma 1: Consider two r.v.'s $\zeta > 0$, and $\eta \ge 0$ over the probabilistic space $(\Omega, \mathfrak{F}, \mathfrak{p})$. Suppose that ζ is measurable w.r. to $\mathcal{G}\subseteq\mathfrak{F}$ and, for a given $a>0$, $\mathfrak{p}(\eta\ge a\mid\mathcal{G})\ge p$. Then: $E[(\zeta+\eta)^{-q}\mid\mathcal{G}]\le p(\zeta+a)^{-q}+(1-p)\zeta^{-q}$, $q\ge 0$. ■

The statements of Proposition 1 and Proposition 2 can be seen as corollaries of a unique key inequality, which will be demonstrated first.

Key inequality: Let ρ be any real number such that

$$\rho\in[\mu,1)\cap((1-K_2)^{1/s},1)\tag{α.1}$$

Letting

$$\xi(n,m):=E\left[\left(\frac{\mu}{\rho}\right)^{(n-m)s}\|P(ns)\|\,\middle|\,\mathcal{P}_1^{ms}\right],\qquad n\ge m$$

the following inequality holds true:

$$\xi(n,m)\le\frac{\rho^{-s}(1-K_2)K_1\xi(n-1,m)+\rho^{-s}}{K_1\xi(n-1,m)+1}\xi(n-1,m),\qquad \forall n\ge m+1\ge 1.$$

Proof: In view of the very definition of $\psi(t)$ one has

$$\|\psi(t)\|\ge\frac{\|\varphi(t)\|}{(1+H^{-2}\|\varphi(t)\|^2)^{1/2}}.$$

Consequently, the persistent excitation Assumption A.3 entails that

$$\mathfrak{p}\left(\lambda_{\min}\left\{\sum_{i=rs+1}^{(r+1)s}\psi(i)\psi(i)'\right\}\ge K_1\mid\mathcal{P}_1^{rs}\right)\ge K_2,\qquad\forall r\ge 0.\tag{α.2}$$

Using (1.3), and recalling that $\mu<1$, one can easily show that $(n\ge m+1)$

$$
\begin{aligned}
&(\mu/\rho)^{(n-m)s}\|P(ns)\|\\
&\le\left(\frac{\mu}{\rho}\right)^{(n-1-m)s}\rho^{-s}\Bigg(\lambda_{\min}[P((n-1)s)^{-1}]\\
&\quad+\lambda_{\min}\left[\sum_{i=(n-1)s+1}^{ns}\psi(i)\psi(i)'\right]\Bigg)^{-1}.
\end{aligned}
\tag{α.3}
$$

By taking conditional expectation w.r. to $\mathcal{P}_1^{(n-1)s}$ at the two sides of (α.3) and applying Lemma 1 with $q=1$ to the right-hand side with

$$
\begin{aligned}
\zeta&=\lambda_{\min}[P((n-1)s)^{-1}],\\
\eta&=\lambda_{\min}\left[\sum_{i=(n-1)s+1}^{ns}\psi(i)\psi(i)'\right],\\
\mathcal{G}&=\mathcal{P}_1^{(n-1)s}
\end{aligned}
$$

the following inequality follows from (α.2):

$$
\begin{aligned}
&E[(\mu/\rho)^{(n-m)s}\|P(ns)\|\mid\mathcal{P}_1^{(n-1)s}]\\
&\le\left(\frac{\mu}{\rho}\right)^{(n-1-m)s}\rho^{-s}\{K_2(\lambda_{\min}[P((n-1)s)^{-1}]+K_1)^{-1}\\
&\quad+(1-K_2)(\lambda_{\min}[P((n-1)s)^{-1}])^{-1}\}.
\end{aligned}
$$

By taking conditional expectation w.r. to $\mathcal{P}_1^{ms}$, one obtains

$$
\begin{aligned}
\xi(n,m)&\le\rho^{-s}K_2E\left[\frac{(\mu/\rho)^{(n-1-m)s}\|P((n-1)s)\|}{1+K_1\|P((n-1)s)\|}\,\middle|\,\mathcal{P}_1^{ms}\right]\\
&\quad+\rho^{-s}(1-K_2)E[(\mu/\rho)^{(n-1-m)s}\|P((n-1)s)\|\mid\mathcal{P}_1^{ms}].
\end{aligned}
\tag{α.4}
$$

On the other hand

$$
\begin{aligned}
&E\left[\frac{(\mu/\rho)^{(n-1-m)s}\|P((n-1)s)\|}{1+K_1\|P((n-1)s)\|}\,\middle|\,\mathcal{P}_1^{ms}\right]\\
&\le E\left[\frac{(\mu/\rho)^{(n-1-m)s}\|P((n-1)s)\|}{1+K_1(\mu/\rho)^{(n-1-m)s}\|P((n-1)s)\|}\,\middle|\,\mathcal{P}_1^{ms}\right]\\
&\hfill\text{(since }\mu\le\rho)\\
&\le\frac{E[(\mu/\rho)^{(n-1-m)s}\|P((n-1)s)\|\mid\mathcal{P}_1^{ms}]}{1+K_1E[(\mu/\rho)^{(n-1-m)s}\|P((n-1)s)\|\mid\mathcal{P}_1^{ms}]}.\\
&\hfill\text{(using Jensen's inequality)}
\end{aligned}
\tag{α.5}
$$

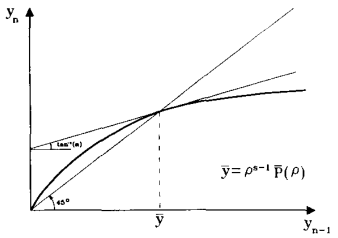

Fig. A.1.

The key inequality follows from (α.4) and (α.5). ■

Under condition (α.1), the diagram of the function $y_n = f(y_{n-1})$ given by the expression

$$y_n = \frac{\rho^{-s}(1-K_2)K_1 y_{n-1} + \rho^{-s}}{K_1 y_{n-1} + 1} y_{n-1}$$

is shown in Fig. A.1.

Then, letting $\overline{P}(\rho) = \rho^{1-s}(1-\rho^s)K_1^{-1}(\rho^s - 1 + K_2)^{-1}$, one can easily conclude from the key inequality that

$$\xi(n, m) \leq \alpha^{n-m}\xi(m, m) + (1-\alpha^{n-m})\rho^{s-1}\overline{P}(\rho), \qquad n \geq m \qquad (\alpha.6)$$

where $\alpha \in (0, 1)$ is a deterministic constant.

Proposition 1) By taking $\rho = \nu$, inequality (α.6) gives:

$$E\left[\left(\frac{\mu}{\nu}\right)^{(n-m)s} \|P(ns)\| \mathcal{P}_1^{ms}\right] \leq \alpha^{n-m}\|P(ms)\| + (1-\alpha^{n-m})\nu^{s-1}\overline{P}(\nu), \qquad n \geq m.$$

The thesis easily follows observing that $P(t) \leq P(t-1)/\mu$. (see (1.2b)).

Proposition 2) By taking $\rho = \mu$ and $m = 0$, from inequality (α.6) it follows that:

$$E[\|P(ns)\|] \leq \max\{\|P(0)\|, \mu^{s-1}\overline{P}(\mu)\}.$$

Bearing in mind again that $P(t) \leq P(t-1)/\mu$, the statement of Proposition 2 is a straightforward consequence of this inequality.

References

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

[2] T. Söderström and P. Stoica, *System Identification.* Englewood Cliffs, NJ: Prentice-Hall, 1989.

[3] R. M. Johnstone, C. R. Johnson, R. R. Bitmead, and B. D. O. Anderson, "Exponential convergence of recursive least squares with exponential forgetting factor," *Syst. Contr. Lett.*, vol. 2, pp. 77–82, 1982.

[4] D. Bertin, S. Bittanti and P. Bolzern, "Tracking of nonstationary systems by means of different prediction error directional forgetting techniques," in *Proc. 2nd Workshop Adaptive Syst. Contr. Sig. Process. Lund.*, 1986.

[5] R. M. Canetti and M. D. España, "Convergence analysis of the least squares identification algorithm with a variable forgetting factor for time-varying linear systems," *Automatica*, vol. 4, pp. 609–612, 1989.

[6] W. S. Krasker and E. Welsch, "Efficient bounded-influence regression estimation," *J. Amer. Stat. Assoc.*, vol. 77, pp. 595–604, 1982.

[7] L. Ljung and S. Gunnarson, "Adaptation and tracking in system identification—A survey," *Automatica*, pp. 7–21, 1990.

[8] A. H. Jazwinski, *Stochastic Processes and Filtering Theory.* New York: Academic, 1970.

[9] H. J. Kushner and A. Shwartz, "Weak convergence and asymptotic properties of adaptive filters with constant gains," *IEEE Trans. Info. Theory*, vol. 30, no. 2, pp. 177–182, 1984.

[10] A. Benveniste and G. Ruget, "A measure of the tracking capability of recursive stochastic algorithms with constant gains," *IEEE Trans. Automat. Contr.*, vol. 27, no. 3, pp. 639–649, 1982.

[11] A. Benveniste, "Design of adaptive algorithms for the tracking of time-varying systems," *Inter. J. Adaptive Contr. Sig. Processing*, vol. 1, pp. 3–29, 1987.

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

[13] L. Ljung, "Optimal and ad hoc adaptation mechanisms," in *Proc. 1st European Contr. Conf.*, Grenoble, 1991, pp. 2013–2020.

[14] S. Bittanti and M. Campi, "Tuning the forgetting factor in RLS identification algorithms," in *Proc. 30th Conf. Decision Contr.*, Brighton, 1991, pp. 1688–1689.

[15] L. Guo, "Estimating time-varying parameters by the Kalman filter based algorithm: Stability and convergence," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 141–147, 1990.

[16] S. Bittanti, M. Campi, and F. Lorito, "Effective identification algorithms for adaptive control," *Inter. J. Adaptive Contr. Sig. Processing*, vol. 6, pp. 221–235, 1992.

Approximate Solution of Large Sparse Lyapunov Equations

Thorkell Gudmundsson and Alan J. Laub

***Abstract*—This note describes a simple method for efficiently estimating the dominant eigenvalues and eigenvectors of the solution to a Lyapunov equation, without first solving the equation explicitly. The method is based on the power method and matrix-vector multiplications and is particularly suitable for problems where those multiplications can be done efficiently, such as where the coefficient matrices are large and sparse or low-rank. The same idea is directly applicable to balanced-truncation order reduction of linear systems.**

I. Introduction

The Lyapunov equation

$$AX + XA^T + BB^T = 0 \qquad (1)$$

(with $A \in \mathbb{R}^{n\times n}$ stable, $X \in \mathbb{R}^{n\times n}$, and $B \in \mathbb{R}^{n\times m}$) is very important in many control applications. Some examples are stability analysis of nonlinear systems [16], [14], optimal $\mathcal{H}_\infty$ control design of linear systems [5], [3], iterative solution of the Riccati equation [15], [21], balancing of linear systems [17], [18], and model reduction methods based on balancing [18], [7].

When the order of the equation is moderate (say, $n < 100$), it can be solved efficiently via the Bartels–Stewart [2] or Hammarling algorithms [10]. These algorithms require O(n^3) computational effort, however, so when the order increases significantly, the need for less computationally intensive algorithms becomes more pronounced. Moreover, when n is large, A is typically sparse and B is of low rank (m is small), while the solution X is, in general, dense. Thus, solving the Lyapunov equation in this case is not only time consuming, but

Manuscript received January 27, 1993; revised May 17, 1993. This research was supported in part by the National Science Foundation under Grant ECS-9120643 and the Air Force Office of Scientific Research under Grant F49620-94-1-0104DEF.

The authors are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 USA.

IEEE Log Number 9400342.

0018-9286/94$04.00 © 1994 IEEE