INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. **12**, 579–597 (1998)

EXPONENTIAL CONVERGENCE OF PRODUCTS OF RANDOM MATRICES: APPLICATION TO ADAPTIVE ALGORITHMS

GEORGE V. MOUSTAKIDES*,†

Department of Computer Engineering and Informatics, University of Patras, 26500 Patras, Greece
Computer Technology Institute (CTI) of Patras, PO Box 1122, 21100 Patras, Greece

SUMMARY

We introduce a novel methodology for analysing well known classes of adaptive algorithms. Combining recent developments concerning geometric ergodicity of stationary Markov processes and long existing results from the theory of Perturbations of Linear Operators we first study the behaviour and convergence properties of a class of products of random matrices, this is turn allows for the analysis of the first and second order statistics of adaptive algorithms without the need of any restrictive conditions imposed on the data (as essential boundedness). Efficient estimates of the convergence rate of adaptive algorithms during the initial transient phase are also presented. These estimates do not rely on the unrealistic Independence Assumption as it is commonly the case in existing literature. © 1998 John Wiley & Sons, Ltd.

Key words: adaptive algorithms; Exponential stability of algorithms; Stochastic approximation

1. INTRODUCTION

Adaptive algorithms are used in many application areas such as filtering, control, communications, biomedical signals processing, etc. Their widespread use in mainly due to their ability to adapt in unknown and changing environments. This practically important characteristics has led to the development of a considerable number of adaptive algorithms each with its own merits and drawbacks.

This paper was recommended for publication by editor P. A. Regalia

*Correspondence to: G. V. Moustakides, Department of Computer Engineering and Informatics, University of Patras, 26500 Patras, Greece.
†This work was completed while the author was Visiting Fellow at the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544.
Contract/grant sponsor: Greek General Secretariat for Research and Technology.
Contract/grant number: ΠΕΝΕΔ95:1584

Received 7 January 1998
Revised 9 June 1998
Accepted 4 September 1998

CCC 0890–6327/98/070579–19 $17·50

The literature dealing with the analysis of adaptive algorithms is substancial and the corresponding approaches can be divided into two main categories, namely the *Deterministic* and the *Stochastic*. As far as the Stochastic approach is concerned most important results are based on the *Stochastic Approximation* theory which was introduced by Ljung with the ODE method[1] and consequently refined in a number of publications.[2-10] These publications primarily focus on the steady state behaviour and asymptotic stability of adaptive algorithms. Recently however there has been a considerable effort to extent the existing theory to also cover transient phenomena for specific algorithms or commonly used algorithmic classes.[11-17]

In order to be more specific let us introduce the problem of interest and the class of algorithms we like to study. Consider the following regression model:

$$y_n = X_n^t W_n + w_n \tag{1}$$

where X_n in the input signal, y_n the desired output, w_n the additive noise and W_n an unknown time varying vector that we like to estimate at every time step n. For simplicity we are going to consider only the scalar case, therefore y_n, w_n are scalars and X_n, W_n are vectors of length N.

A rich class of algorithms used to estimate W_n can be defined by the following general recursion:

$$\hat{W}_{n+1} = \hat{W}_n + \mu(y_n - X_n^t \hat{W}_n) Z_n \tag{2}$$

where $\hat{W}_n$ is the estimate of W_n at time n, Z_n is a vector of length N known as *gain* or *regression vector* and μ is a positive scalar known as *step size*. By proper selection of the vector Z_n one can obtain most well known adaptive algorithms used in practice. Examples are the LMS, Normalized LMS, Signed Regressor LMS, RLS, Kalman filter based algorithms, Newton type algorithms, etc.

A common characteristic encountered in all algorithms covered by the model in (2) is that their analysis can be reduced to the analysis of a certain product of random matrices. Existing results attempt to describe the behaviour of the statistics of such products and establish some form of exponential convergence. However, these publications base their analysis on very stringent conditions resulting in the exclusion of common combinations of algorithms and data types. Characteristic example is Condition i. in Theorem 3.2 in Reference 12 which is not valid even for LMS and Gaussian i.i.d. data. In fact, a publication appeared very recently,[15] by the same authors, focusing exclusively on LMS and introducing conditions that allow for Gaussian data. Unfortunately even this new set of conditions is not satisfied by common data types. Specifically if the input data are i.i.d. and polynomial transformations of Gaussian random variables then Condition (6) in Reference 15 fails. This of course suggests that even for the simple case of LMS the conditions imposed in Reference 15 continue to be restrictive. It should be noted however that for the special cases of RLS and Kalman Filter algorithms these stringent conditions can be significantly relaxed.[11,13,14] Furthermore in Reference 15, Remark 3, it is mentioned that for general algorithms it is possible to relax these stringent conditions when one is limited to the stationary case.

In this paper we primarily focus on products of random matrices appearing in the analysis of the adaptive algorithms defined by (2). More specifically we propose a novel approach for studying the behaviour of these products that leads to efficient estimates of their performance. Our methodology is based on a proper combination of the theory of Markov processes[18] (specifically results concerning geometric ergodicity) and the theory of Perturbations of Linear Operators in Banach spaces.[19] The results we develop constitute our main tool for analyzing the

first and second order statistics of the estimate $\hat{W}_n$ of the adaptive algorithms in (2). It is worth noting that our analysis also yields efficient estimates for the convergence rate of the adaptive algorithms during the initial transient phase without making use of any unrealistic assumptions (as the Independence Assumption). Furthermore, as far as LMS is concerned, our results are applicable to data that can be ANY polynomial (and even exponential) transformation of Gaussian random variables.

2. ANALYSIS OF PRODUCTS OF RANDOM MATRICES

As was stated in the Introduction, our primary concern in the study of the behaviour of products of a specific class of random matrices. The matrices we like to consider will be comprised of elements that are memoryless non-linear transformations of a vector *Markov process* $\{\xi_n\}$. Thus let us assume that a stationary discrete time Markov process $\{\xi_n\}$ exists evolving of some state space X equipped with a σ-field $\mathscr{B}(X)$. The products that appear in the analysis of algorithms of the form of (2) and are consequently to our interest are

$$\mathbf{U}_n(\mu) = \left[\mathbf{I}_K + \sum_{l=1}^{r} \mu^l T_l(\xi_1)\right] \cdots \left[\mathbf{I}_K + \sum_{l=1}^{r} \mu^l T_l(\xi_n)\right] \tag{3}$$

with $\mathbf{T}_l(\xi)$, $l = 1, \ldots, r$, matrix functions of ξ of dimensions $K \times K$, $\mathbf{I}_K$ the identity matrix of the same dimensions and $\mu > 0$ a scalar variable that is to be assumed 'small' (i.e. $\mu \ll 1$) and corresponding to the step size μ of (2).

2.1. Assumptions and definitions

Let us first introduce our assumptions elaborating also on their meaning and generality. As we have seen, a key characteristic of the random matrices is the existence of the Markov vector process $\{\xi_n\}$ which controls them. To stationary Markov processes we usually assign two types of expectation, first is the steady state expectation which we denote by $\mathbb{E}\{\cdot\}$ and second the conditional expectation given that state $\xi_0 = \xi$ which we denote by $\mathbb{E}_\xi\{\cdot\}$. We then have the following assumptions:

$\mathscr{A}1$. The Markov process $\{\xi_n\}$ is stationary and *geometrically ergodic* in the following sense. We can find a scalar measurable function $V(\xi) \geqslant 1$ (*drift function*) such that (a) the steady state expectation $\mathbb{E}\{V(\xi_n)\}$ is finite and (b) for all measurable functions $g(\xi)$ with $|g(\xi)| \leqslant V(\xi)$ there exist constants $\rho \in [0, 1)$ and R both independent of $g(\xi)$ such that

$$\sup_{|g| \leqslant V} \sup_{\xi} \frac{|\mathbb{E}_\xi\{g(\xi_n)\} - \mathbb{E}\{g(\xi_n)\}|}{V(\xi)} \leqslant R\rho^n \tag{4}$$

$\mathscr{A}2$. The matrices $\mathbf{T}_l(\xi)$ entering in (3) satisfy

$$\mathbb{E}_\xi\{\|\mathbf{T}_l(\xi_1)\| V(\xi_1)\} \leqslant cV(\xi), \quad l = 1, \ldots, r \tag{5}$$

where $c < \infty$ is a constant, $V(\xi)$ is a drift function as in $\mathscr{A}1$ and the norm of a matrix $\mathbf{C}$ is defined as $\|\mathbf{C}\|^2 = \text{trace}\{\mathbf{C}^t\mathbf{C}\}$ (Frobenius norm).

The first assumption, with the help of the drift function $V(\xi)$, introduces a form of (*geometric*) ergodicity for the Markov process $\{\xi_n\}$. Condition (4) is of principal significance to our analysis. It basically suggests that the conditional expectation of $g(\xi_n)$, given that the initial state is ξ, tends with an exponential like speed to the steady state expectation. This convergence is not necessarily uniform with respect to the initial state ξ (since $V(\xi)$ can be unbounded) but, on the other hand, it is uniformly controlled through the drift function $V(\xi)$ in the sense that $|\mathbb{E}_\xi\{g(\xi_n)\} - \mathbb{E}\{g(\xi_n)\}| \leqslant \rho^n R V(\xi)$ for all functions satisfying $|g(\xi)| \leqslant V(\xi)$.

Assumption $\mathscr{A}1$ constitutes a generalization to the corresponding notion of geometric ergodicity encountered in finite state Markov chains[18] (where this convergence is uniform). It is clear that the introduction of the drift function $V(\xi)$, having the specified properties, seems rather arbitrary. It should be noted however that it is possible to guarantee its existence through a set of more reasonable assumptions imposed directly on the transition probability of $\{\xi_n\}$.[18] Unfortunately such an approach turns out to be very complicated requiring the definition of various unnecessary (to our goal) notions and the introduction of a number of intermediate results before obtaining the desired condition specified in $\mathscr{A}1$. Therefore we decided to follow this seemingly arbitrary but considerably simpler approach. Assumptions imposed on conditional probabilities leading to $\mathscr{A}1$ or other easily verifiable sufficient conditions guaranteeing $\mathscr{A}1$ can be found in Reference 18. Such sufficient conditions will also be presented in Section 3.

Assumption $\mathscr{A}2$ introduces technical constraints on the matrices entering in the product we like to study. These constraints are trivially satisfied if the matrices $\mathbf{T}_l(\xi)$ are essentially bounded. This is for example the case in certain normalized versions of adaptive algorithms or if the data are essentially bounded. It should also be noted that the richness of the family of allowable matrices satisfying (5) strongly depends on the form of the drift function $V(\xi)$. In Section 3 we are going to see that these families turn out to be sufficiently rich for most commonly used data processes.

Let us now introduce a useful space of vector functions along with a class of linear operators defined in this space. Let from now on $V(\xi) \geqslant 1$ denote a drift function satisfying $\mathscr{A}1$. Consider the class $\mathscr{L}_V^\infty$ of all measurable vector functions $G(\xi)$ of length K that satisfy $\sup_\xi(\|G(\xi)\|/V(\xi)) < \infty$. The space $\mathscr{L}_V^\infty$ equipped with the norm

$$\|G(\xi)\|_V = \sup_\xi \frac{\|G(\xi)\|}{V(\xi)} \tag{6}$$

constitutes a Banach space. In a sense $\mathscr{L}_V^\infty$ contains all vector functions for which we can *a priori* guarantee existence of their steady state expectation along with exponential ergodicity for their elements in the form described by Assumption $\mathscr{A}1$.

Consider now linear operators mapping $\mathscr{L}_V^\infty$ to $\mathscr{L}_V^\infty$, If $\mathscr{M}$ is such an operator then its norm induced by the corresponding norm in $\mathscr{L}_V^\infty$ can be defined as

$$\|\mathscr{M}\|_V = \sup_{\|G\|_V \leqslant 1} \|\mathscr{M}G(\xi)\|_V \tag{7}$$

For any $G(\xi) \in \mathscr{L}_V^\infty$ let us now introduce the expectation

$$Q_n(\mu, \xi) = \mathbb{E}_\xi\{\mathbf{U}_n(\mu)G(\xi_n)\} = \mathbb{E}_\xi\left\{\left[\prod_{i=1}^{n}[\mathbf{I}_K + \sum_{l=1}^{r}\mu^l\mathbf{T}_l(\xi_i)]\right]G(\xi_n)\right\} \tag{8}$$

which constitutes the main subject of this section. If we define the linear operator $\mathscr{T}(\mu)$ as follows

$$\mathscr{T}(\mu)G(\xi) = \mathbb{E}_\xi\left\{\left[\mathbf{I}_K + \sum_{l=1}^{r} \mu^l \mathbf{T}_l(\xi_1)\right] G(\xi_1)\right\} \tag{9}$$

then using stationarity and the Markov property for conditional expectations we can see that $Q_n(\mu, \xi)$ can be written as

$$Q_n(\mu, \xi) = \mathscr{T}(\mu)^n G(\xi) \tag{10}$$

In other words $Q_n(\mu, \xi)$ is the result of the n-times repeated application of $\mathscr{T}(\mu)$ on the element $G(\xi)$ of $\mathscr{L}_V^\infty$. To write $\mathscr{T}(\mu)$ using the notation in Reference 19 (which we are going to reference several times in the sequel) let us define the following linear operators:

$$\mathscr{T}G(\xi) = \mathbb{E}_\xi\{G(\xi_1)\}, \mathscr{T}^{(l)}G(\xi) = \mathbb{E}_\xi\{\mathbf{T}_l(\xi_1)G(\xi_1)\}, \quad l = 1, \ldots, r \tag{11}$$

then $\mathscr{T}(\mu)$ can also be written as

$$\mathscr{T}(\mu) = \mathscr{T} + \mu\mathscr{T}^{(1)} + \cdots + \mu^r\mathscr{T}^{(r)} \tag{12}$$

Although the matrices $\mathbf{T}_l(\xi)$ have finite dimensions, the corresponding operator $\mathscr{T}(\mu)$, defined in (9), is in general infinite dimensional. This is in fact the main source of difficulty in the analysis. The next subsection will be devoted to a detailed study of the behaviour of $\mathscr{T}(\mu)^n$ by identifying characteristics that are important to the analysis of adaptive algorithms.

2.2. *Decomposition of* $\mathscr{T}(\mu)^n$

Finite dimensional operators are known to be decomposable into their eigenprojections and so are their powers. With the next theorem we will see that, in a sense, this is also possible for the infinite dimensional operator $\mathscr{T}(\mu)$.

Theorem 1

Let assumptions $\mathscr{A}1$, $\mathscr{A}2$ be valid then, for small enough μ, the operator $\mathscr{T}(\mu)$ can be decomposed as $\mathscr{T}(\mu) = \mathscr{F}(\mu) + \mathscr{H}(\mu)$ with the characteristic $\mathscr{T}(\mu)^n = \mathscr{F}(\mu)^n + \mathscr{H}(\mu)^n$. For the operator $\mathscr{H}(\mu)$ we can find constants $c > 0$ and $1 > \bar{\rho} \geqslant 0$ such that

$$\| \mathscr{H}(\mu)^n \|_V \leqslant c\bar{\rho}^n \tag{13}$$

For the operator $\mathscr{F}(\mu)$ we have that it is finite dimensional and of dimension K and if $f_i(\mu)$, $\mathscr{P}_i(\mu)$, $\mathscr{D}_i(\mu) = [\mathscr{F}(\mu) - f_i(\mu)1]\mathscr{P}_i(\mu) = [\mathscr{T}(\mu) - f_i(\mu)1]\mathscr{P}_i(\mu)$, $i = 1, \ldots, s$, are its eigenvalues and its corresponding eigenprojections and eigennilpotents, then the following approximations apply:

$$f_i(\mu) = 1 + \mu\lambda_i + \mu o(1) \tag{14}$$

$$\mathscr{P}_i(\mu) = \mathbf{P}_i\mathscr{P} + \mu O_V(1) \tag{15}$$

$$\mathscr{D}_i(\mu)^k = \mu^k\mathbf{D}_i^k\mathscr{P} + \mu^k o_V(1) \tag{16}$$

where if $\mathbf{T}$ denotes the constant matrix $\mathbf{T} = \mathbb{E}\{\mathbf{T}_1(\xi_1)\}$ then λ_i, $\mathbf{P}_i$, $\mathbf{D}_i = (\mathbf{T} - \lambda_i\mathbf{I})\mathbf{P}_i$, $i = 1, \ldots, s$, denote its eigenvalues and its corresponding eigenprojections and eigennilpotents, $\mathscr{P}$ denotes the operator induced by the steady state expectation (i.e. $\mathscr{P} = \mathbb{E}\{\cdot\}$), $o(1)$ a scalar quantity that

satisfies $\lim_{\mu\to 0} o(1) = 0$, $o_V(1)$ an operator that satisfies $\lim_{\mu\to 0} \| o_V(1) \|_V = 0$ and $O_V(1)$ an operator that has uniformly, in μ, bounded $\| \cdot \|_V$ norm.

Proof. The proof can be found in Appendix A. □

Since $\mathscr{F}(\mu)$ is finite dimensional we can further decompose its power $\mathscr{F}(\mu)^n$ using known results from finite dimensional operator theory. Specifically by selecting $\phi(\zeta) = \zeta^n$ in equations (5.50) and (5.51) Reference of 19, p. 45 we obtain

$$\mathscr{F}(\mu)^n = \sum_{i=1}^{s} f_i(\mu)^n P_i(\mu) + f_i(\mu)^{n-1} \binom{n}{1} \mathscr{D}_i(\mu) + \cdots + f_i(\mu)^{n-m_i+1} \binom{n}{m_i - 1} \mathscr{D}_i(\mu)^{m_i-1} \tag{17}$$

where m_i, $i = 1, \ldots, s$ are the multiplicities of the corresponding eigenvalues. Using now Theorem 1 and (17) we can introduce the following theorem that establishes efficient approximations for $\mathscr{T}(\mu)^n$.

Theorem 2

Let the assumptions of Theorem 1 be valid and the matrix $\mathbf{T} = \mathbb{E}\{\mathbf{T}_1(\xi_1)\}$ have eigenvalues with strictly negative real parts. Define the constant matrix $\mathbf{F}(\mu) = \mathbf{I}_K + \mu\mathbf{T} + \mu^2\mathbf{T}'$ with $\mathbf{T}'$ any arbitrary constant matrix. Let $\mathscr{P} = \mathbb{E}\{\cdot\}$ denote the operator induced by the steady state expectation then, for small enough μ, we can find constants $\alpha > 0$ and $1 > \bar{\rho} \geqslant 0$ such that

$$\mathscr{T}(\mu)^n = \mathbf{F}(\mu)^n \mathscr{P} + (1 - \mu\alpha)^n o_V(1) + \bar{\rho}^n O_V(1) \tag{18}$$

$$\mathscr{P}\mathscr{T}(\mu)^n = \mathbf{F}(\mu)^n \mathscr{P} + (1 - \mu\alpha)^n \mathscr{P} o_V(1) + \mu\bar{\rho}^n \mathscr{P} O_V(1) \tag{19}$$

where $O_V(1)$ denotes an operator that has, for small enough μ, a uniformly bounded norm in n and μ and $o_V(1)$ an operator that satisfies, uniformly in n, $\lim_{\mu\to 0} \| o_V(1) \|_V = 0$.

Proof. The proof can be found in Appendix A. □

Comments. Relations (18), (19) constitute the basis for obtaining estimates for the statistics of adaptive algorithms as in References 12 and 13. Notice that their main characteristic is that the repeated application of the operator $\mathscr{T}(\mu)$ on an element $G(\xi) \in \mathscr{L}_V^\infty$ resembles, in a sense, the behaviour of the constant matrix $\mathbf{F}(\mu)$ applied repeatedly on the constant vector $\mathscr{P}G(\xi) = \mathbb{E}\{G(\xi_1)\}$. It is clear that (19) is a refinement of (18) for the case where we consider expectation (instead of conditional expectation) of products of random matrices.

A last result in this section consists in observing that the spectral radius of the operator $\mathscr{T}(\mu)$ can be well approximated through the eigenvalues of the constant matrix $\mathbf{T} = \mathbb{E}\{\mathbf{T}_1(\xi_1)\}$.

Corollary 1

Under the assumptions of Theorem 1 we have that the spectral radius of $\mathscr{T}(\mu)$ can be written as

$$\lim_{n\to\infty} \sqrt[n]{\| \mathscr{T}(\mu)^n \|_V} = 1 + \mu \max_i \operatorname{Re}(\lambda_i) + \mu o(1) \tag{20}$$

where λ_i, $i = 1, \ldots, s$, are the eigenvalues of the matrix $\mathbf{T} = \mathbb{E}\{\mathbf{T}_1(\xi_1)\}$ and $\mathrm{Re}(\cdot)$ denotes the real part.

Proof. From Theorem 1 we have $\mathscr{T}(\mu)^n = \mathscr{F}(\mu)^n + \mathscr{H}(\mu)^n$ with $\| \mathscr{H}(\mu)^n \|_V \leqslant c\bar{\rho}^n$ and $0 \leqslant \bar{\rho} < 1$. On the other hand, from (14) we have that the eigenvalues of the finite dimensional operator $\mathscr{F}(\mu)$, for small enough μ, can all have magnitude larger than $\bar{\rho}$, consequently because of (17) we can write

$$\lim_{n \to \infty} \sqrt[n]{\| \mathscr{T}(\mu)^n \|_V} = \lim_{n \to \infty} \sqrt[n]{\| \mathscr{F}(\mu)^n \|_V} \tag{21}$$

Since $\mathscr{F}(\mu)$ is finite dimensional its spectral radius is equal to the maximum amplitude of its eigenvalues, i.e. $\max_i\{|f_i(\mu)|\}$ which, because of (14), is equal to the desired expression. □

3. EXAMPLES

In this section we present characteristic examples of Markov processes and families of matrix functions $\mathbf{T}_l(\xi)$ that satisfy our assumptions and therefore can be studied using the theory we developed in the previous section. Specifically for several well known Markov processes we are going to explicitly identify drift functions that satisfy Assumption $\mathscr{A}1$ and also families of matrix functions $\mathbf{T}_l(\xi)$ that satisfy $\mathscr{A}2$.

3.1. Linear state space model

Consider the Markov process generated by the following mechanism:

$$\xi_n = \mathbf{C}\xi_{n-1} + \mathbf{D}\eta_n \tag{22}$$

where $\{\eta_n\}$ is a white noise vector process and $\mathbf{C}$, $\mathbf{D}$ are constant matrices. This model includes the case of AR and ARMA processes which are the most common data models used in practice.

To identify a drift function $V(\xi)$ we apply Theorem 16.0.1 (Reference 18, p. 383). In particular, using techniques similar to Theorem 16.5.1 (Reference 18, p. 404), we have that if $\mathbf{C}$ has all its eigenvalues strictly inside the unit circle, the pair $(\mathbf{C}, \mathbf{D})$ is controllable, and η_n has an everywhere positive density then a function $V(\xi) \geqslant 1$ is a drift function (i.e. satisfies Condition $\mathscr{A}1$) if it satisfies the following inequality

$$\mathbb{E}_\xi\{V(\xi_1)\} \leqslant \nu V(\xi) + L\mathbb{1}_A(\xi) \tag{23}$$

where $0 \leqslant \nu < 1$ and L are constants and A is a compact set with $\mathbb{1}_A(\xi)$ denoting its indicator function. Inequality (23) is known as *geometric drift condition* (Reference 18, p. 367).

It turns out that there is no unique solution to (23). To find a specific one, let $\mathbf{S}$ be the solution to the Lyapunov equation $\mathbf{S} = \mathbf{C}^{\mathrm{t}}\mathbf{S}\mathbf{C} + \mathbf{I}_K$, then if $\lambda_{\max}(\mathbf{S})$ denotes the maximum eigenvalue of $\mathbf{S}$ we have $\lambda_{\max}(\mathbf{S}) \geqslant 1$ and $\mathbf{S} \geqslant \mathbf{I}_K \geqslant \lambda_{\max}^{-1}(\mathbf{S})\mathbf{S}$. Define $\kappa = 1 - \lambda_{\max}^{-1}(\mathbf{S})$ then $0 \leqslant \kappa < 1$ and $\kappa\mathbf{S} \geqslant \mathbf{S} - \mathbf{I}_K = \mathbf{C}^{\mathrm{t}}\mathbf{S}\mathbf{C}$. Using (22) and the fact that $(\xi^{\mathrm{t}}\mathbf{S}\xi)^{1/2}$ is a vector norm we have from (22) and for $n = 1$

$$(\xi_1^{\mathrm{t}}\mathbf{S}\xi_1)^{1/2} \leqslant (\xi^{\mathrm{t}}\mathbf{C}^{\mathrm{t}}\mathbf{S}\mathbf{C}\xi)^{1/2} + (\eta_1^{\mathrm{t}}\mathbf{D}^{\mathrm{t}}\mathbf{S}\mathbf{D}\eta_1)^{1/2} \leqslant \kappa^{1/2}(\xi^{\mathrm{t}}\mathbf{S}\xi)^{1/2} + \lambda_{\max}^{1/2}(\mathbf{D}^{\mathrm{t}}\mathbf{S}\mathbf{D})\,\|\eta_1\| \tag{24}$$

If η_n has up to pth order bounded moments ($\infty > p \geqslant 1$) then $V(\xi) = 1 + (\xi^t \mathbf{S} \xi)^{p/2}$ can be shown to be a drift function. Indeed with the help of Hölder's inequality applied to (24), with q satisfying $(1/p) + (1/q) = 1$ and $1 > \varepsilon > 0$ we have that

$$(\xi_1 \mathbf{S} \xi_1)^{p/2} \leqslant (\kappa^{q/2} + \varepsilon^{q/2})^{p/q} \left((\xi \mathbf{S} \xi)^{p/2} + \frac{[\lambda^{\max}(\mathbf{D}^t \mathbf{S} \mathbf{D})]^{p/2}}{\varepsilon^p} \| \eta_1 \|^p \right) \tag{25}$$

Taking expectation with respect to η_1 and selecting ε small enough it is clear that we can find constants $0 < v' < 1$ and L' such that $\mathbb{E}_\xi \{V(\xi_1)\} \leqslant v' V(\xi) + L'$. Then for any v with $v' < v < 1$ we have that the drift condition is satisfied with $L = L'/(v - v')$ and $A = \{\xi : V(\xi) \leqslant L\}$.

With the drift function just defined we can ensure ergodicity of the Markov process for non-linear functions that can be bounded by a polynomial of order p. A significantly more interesting situation occurs when η_1 has all its moments bounded in the sense that $\mathbb{E}\{e^{\delta \| \eta_1 \|}\} < \infty$ for some $\delta > 0$. Here we can define a drift function of the form $V(\xi) = e^{\delta' (\xi^t \mathbf{S} \xi)^{1/2}}$ with $\delta' > 0$. The proof is analogous to the previous case consequently we do not present any further details. The existence of exponential like drift functions is very desirable because, as we are going to see next, it enriches significantly the class of allowable matrix functions $\mathbf{T}_l(\xi)$ that satisfy Assumption $\mathscr{A}2$.

Let us now consider the problem of satisfying Assumption $\mathscr{A}2$. In the case of a polynomial $V(\xi)$, if we make no further assumptions, we can assume validity of (5) only for matrix functions $\mathbf{T}_l(\xi)$ that are essentially bounded. If on the other hand we use the exponential drift function then we can easily show that there exists $\delta'' > 0$ such that for any $\mathbf{T}_l(\xi)$ with $\sup_\xi \| \mathbf{T}_l(\xi) \| / (e^{\delta'' (\xi^t \mathbf{S} \xi)^{1/2}}) < \infty$, condition (5) in $\mathscr{A}2$ is satisfied. This last inequality clearly defines a very rich class of allowable matrices $\mathbf{T}_l(\xi)$ because it includes matrices whose elements can be ANY polynomial function of ξ (or function that can be bounded by polynomial) and also functions that can grow exponentially as $e^{a \| \xi \|}$ for small enough a.

Depending on the way the moments of η_n grow we can define even more general drift functions. If for instance we have for some $\delta > 0$ and $p \geqslant 1$ that $\mathbb{E}\{e^{\delta \| \eta_1 \|^p}\} \leqslant c < \infty$ (as in the Gaussian case where this condition is valid for $1 \leqslant p \leqslant 2$), we can define as above drift functions of the form $V(\xi) = e^{\delta' (\xi^t \mathbf{S} \xi)^{p/2}}$ and a similar bound for the matrices $\mathbf{T}_l(\xi)$ so that our two assumptions are valid. If $p > 1$ then in addition to any polynomial-like function we can also allow functions for the elements of $\mathbf{T}_l(\xi)$ that can grow exponentially with $\| \xi \|$ as $e^{a \| \xi \|}$ without any constraint on the value of a.

One might argue that the condition $\mathbb{E}\{e^{\delta \| \eta_1 \|^p}\} < \infty$, used above, is similar to the corresponding conditions defined in References 12 and 15 (i.e. Condition (i), Theorem 3.2 in Reference 13 and Condition (6) in Reference 15). The latter conditions however, translated into our terminology, imply the relation $\mathbb{E}\{e^{\delta \| \mathbf{T}_l(\xi_1) \|}\} < \infty$ which is considerably more stringent to satisfy. Characteristic example constitutes the Gaussian $\{\eta_n\}$ case where all our previous analysis applies with $p = 2$ to matrices $\mathbf{T}_l(\xi)$ with elements that are polynomial and even exponential transformations of ξ whereas $\mathbb{E}\{e^{\delta \| \mathbf{T}_l(\xi_1) \|}\} < \infty$ can be true only when $\| \mathbf{T}_l(\xi) \|$ is at most quadratic with ξ.

3.2. Non-linear state space model

The foregoing results can be extended to non-linear state space models of the form

$$\xi_n = \mathbf{C} \xi_{n-1} + \Phi(\xi_{n-1}) + \mathbf{D} \eta_n \tag{26}$$

with $\{\eta_n\}$ i.i.d. and $\Phi(\xi)$ such that $\sup_\xi \| \Phi(\xi) \| / (1 + \| \xi \|) < \infty$ and $\lim_{\| \xi \| \to \infty} \| \Phi(\xi) \| / (1 + \| \xi \|) = 0$ uniformly in $\| \xi \|$. If again η_n has an everywhere positive distribution and the pair

(**C**, **D**) is controllable then we can define the same drift functions as in the linear case with exactly the same properties as far as Assumptions $\mathscr{A}1$ and $\mathscr{A}2$ concerned.

3.3. Sublinear state space model

We have seen in the linear (and non-linear) state space model that in order to have a rich family of matrix functions satisfying condition $\mathscr{A}2$ we need to use exponential like drift function $V(\xi)$. The problem with this type of function is the requirement of existence of all moments of η_n. Such a condition is rather restrictive if it is for example applied to the case where $\{\xi_n\}$ is i.i.d. With the following non-linear model we can overcome this problem. Consider the system

$$\xi_n = \Phi(\xi_{n-1}) + \eta_n \tag{27}$$

where as before η_n is i.i.d. and has an everywhere positive distribution. Let $0 < \delta < 1$ be such that $\sup_\xi \| \Phi(\xi) \| / (1 + \| \xi \|^\delta) < \infty$ (sublinear case) and consider $V(\xi) = 1 + \| \xi \|^p$. We can then show that $V(\xi)$ is a drift function if $\mathbb{E}\{ \| n_n \|^p \} < \infty$. If we also consider matrix functions $\mathbf{T}_l(\xi)$ that satisfy $\sup_\xi \| \mathbf{T}_l(\xi) \| / (\mathbf{1} + \| \xi \|^q) < \infty$ with $q = (1/\delta - 1)p$ and assume that $\mathbb{E}\{ \| \eta_n \|^{p/\delta} \} < \infty$ then, with the help of Minkowski's inequality, we can write

$$\begin{aligned} \mathbb{E}_\xi\{ \| \mathbf{T}_l(\xi_1) \| V(\xi_1) \} &\leqslant c_1 \mathbb{E}_\xi\{1 + \| \xi_1 \|^{p+q}\} \leqslant c_2(1 + \| \Phi(\xi) \|^{p+q}) \\ &\leqslant c_3(1 + \| \xi \|^{\delta(p+q)}) = c_3 V(\xi) \end{aligned} \tag{28}$$

for c_1, c_2, c_3 proper constants, and thus satisfy (5). In other words, for this case we can have a polynomial drift function (requiring only boundedness of a finite number of moments of η_n) and matrices $\mathbf{T}_l(\xi)$ whose norm can also be bounded by a polynomial. Notice now that $\Phi(\xi) = 0$ is definitely sublinear consequently our general analysis when applied to the i.i.d. $\{\xi_n\}$ case does not require any extra constraints as compared to an analysis developed specifically for i.i.d. data.

Comments. We must note that everywhere positivity of the distribution of η_n is not actually necessary. In Reference 18 one can find sufficient conditions that can guarantee the existence of drift functions for more general processes.

The above Markov processes are only a few characteristic examples of processes that can be treated with the theory developed in Reference 18. Other interesting models are presented in Reference 18 and, depending on the case, various drift functions $V(\xi)$ can be constructed. It turns out that, most of the time, it is easy to show existence of polynomial like drift functions.

4. APPLICATION TO ADAPTIVE ALGORITHMS

Let us now turn to the adaptive algorithms defined by equation (2). In order to be able to apply the theory we developed in the previous sections we need to impose certain constraints on the data and the algorithms we like to study.

First we will assume that Assumption $\mathscr{A}1$ continuous to hold. In other words there exists a stationary Markov process $\{\xi_n\}$, not necessarily observable, that controls our data and satisfies the geometric ergodicity condition (4) with the help of a drift function $V(\xi)$. Assumption $\mathscr{A}2$, on the other hand, must be modified as follows.

$\mathscr{A}2'$. The vectors X_n, Z_n are constant non-linear transformations of the Markov process $\{\xi_n\}$ of the form $X_n = X(\xi_n), Z_n = Z(\xi_n)$ satisfying

$$\mathbb{E}_\xi\{\|X(\xi_1)\|^l \|Z(\xi_1)\|^l V(\xi)\} \leqslant cV(\xi), \quad l = 1, 2 \tag{29}$$

Finally we need a last assumption referring to the additive noise w_n and the change mechanism of the vector W_n.

$\mathscr{A}3$. The vector W_n satisfies the following recursion:

$$W_{n+1} = W_n + \gamma U_{n+1} \tag{30}$$

with γ a scalar constant and $\{U_n\}$ a zero mean stationary white noise vector sequence independent of $\{\xi_n\}$ and $\{w_n\}$ with covariance matrix $\mathbf{Q}_U$. The additive noise $\{w_n\}$ is stationary white independent of $\{\xi_n\}$ and $\{U_n\}$ with variance σ_w^2.

The assumption that the data process $\{X_n\}$ is obtained through a constant non-linear transformation of the stationary Markov process $\{\xi_n\}$ is not very restrictive. As we have seen in Section 3 this allows for very rich families of data. Furthermore, in practice, stationarity or at least stationarity within a block of data is a very common assumption.

The most crucial and restrictive assumption is the requirement that Z_n must be a constant vector function of ξ_n. Since in most well known algorithms Z_n is a function of the data history $X_n, \ldots, X_1$ this might seem as if we practically limit ourselves to algorithms where Z_n can only be a function of X_n (as for example LMS). Fortunately this limitation does not apply. The reason is that if $\{\xi_n\}$ is a stationary Markov process so is any finite combination of its states of the form $\Xi_n = [\xi_n^t \cdots \xi_{n-L}^t]^t$. Consequently, if we assume that X_n, Z_n are now functions of Ξ_n instead of ξ_n this allows Z_n to have finite memory. Of course this still excludes algorithms that use exponential windowing on the data (as constant forgetting factor RLS) where Z_n has infinite memory but, on the other hand, covers cases where a sliding window is applied (as sliding window RLS).

Relation (29) in $\mathscr{A}2'$ is technical and corresponds to (5) of $\mathscr{A}2$. The reason that in this constraint parameter l takes upon the values 1 and 2 is because we are going to study the first and second order statistics of $\hat{W}_n$. If one intends to study up to the rth order statistics of $\hat{W}_n$ then l must go up to r.

Finally in $\mathscr{A}3$ we specify a random walk model for the change of W_n and also assume the simplest and most common additive noise model for w_n.

4.1. Convergence in the mean

In this subsection we are going to examine the convergence in the mean of $\hat{W}_n$ towards W_n. If we denote by $\Delta_n = \hat{W}_n - W_n$ the estimation error, we can then write

$$\Delta_n^t = \Delta_0^t \prod_{j=1}^{n} (\mathbf{I}_N - \mu X_j Z_j^t) + \sum_{j=1}^{n} (\mu w_j Z_j^t - \gamma U_{j+1}^t) \prod_{l=j+1}^{n} (\mathbf{I}_N - \mu X_l Z_l^t) \tag{31}$$

The reason we introduced the transposed version of Δ_n is only technical and in order to be consistent with our notation in the previous section. Taking expectation in (31) yields

$$\mathbb{E}\{\Delta_n^t\} = \Delta_0^t \mathbb{E}\left\{\prod_{j=1}^{n} (\mathbf{I}_N - \mu X_j Z_j^t)\right\} \tag{32}$$

which is of the form we analysed in Section 3. We can thus show the following theorem concerning the exponential rate of (mean) convergence.

Theorem 3

Let Assumptions $\mathscr{A}1$, $\mathscr{A}2'$ and $\mathscr{A}3$ be valid then the exponential rate of convergence of the expectation of the estimation error Δ_n satisfies

$$\lim_{n\to\infty} \frac{-\log\{\| \mathbb{E}\{\Delta_n\} \|\}}{n} = \mu \min_i \mathrm{Re}(\lambda_i) + \mu o(1) \tag{33}$$

where λ_i $i = 1, \ldots, s$, are the eigenvalues of the constant matrix $\mathbf{A} = \mathbb{E}\{X_1 Z_1^t\}$ and $\mathrm{Re}(\cdot)$ denotes real part.

Proof. The proof is based on Theorem 1 and it is presented in Appendix B. □

With the above result it is easy to analyse the stability properties of the algorithm during the transient phase. We have the following corollary:

Corollary 2

If all eigenvalues of the matrix $\mathbf{A} = \mathbb{E}\{X_1 Z_1^t\}$ have positive real parts then, for small enough μ, the algorithm is stable in the mean. If on the other hand at least one eigenvalue of $\mathbf{A}$ has negative real part then, for small enough μ, the algorithm is unstable (in the mean).

What is interesting to note is that if we apply the *Independence Assumption* (i.e. that $\hat{W}_n$ is independent from the data) during the transient phase this will also yield as rate of convergence $\mu \min_i \mathrm{Re}\{\lambda_i\} + \mu o(1)$. In other words under this commonly used (but erroneous) assumption we obtained an expression for the rate that is correct up to a first order approximation with respect to μ. Unfortunately this property does not apply to higher order approximations in μ, which are possible (see Reference 19).

4.2. Second order statistics

In this subsection we will focus on estimates for the second order statistics of Δ_n. However, instead of obtaining estimates for the covariance matrix of Δ_n we would like to present a slightly more general result. Thus let us define the sequence of deterministic matrices $\{\mathbf{\Pi}_n\}$ using the following recursion:

$$\mathbf{\Pi}_{n+1} = (\mathbf{I}_N - \mu\mathbf{A})\mathbf{\Pi}_n(\mathbf{I}_N - \mu\mathbf{A})^t + \mu^2\sigma_w^2\mathbf{Q}_Z + \gamma^2\mathbf{Q}_U, \qquad \mathbf{\Pi}_0 = \Delta_0\Delta_0^t \tag{34}$$

where $\mathbf{Q}_Z = \mathbb{E}\{Z_1 Z_1^t\}$, $\mathbf{Q}_X = \mathbb{E}\{X_1 X_1^t\}$, $\mathbf{Q}_U = \mathbb{E}\{U_1 U_1^t\}$ and $\mathbf{A} = \mathbb{E}\{Z_1 X_1^t\}$. Let also $Y_n = Y(\xi_n)$ and $G_n = G(\xi_n)$ be constant vector functions of an underlying Markov process. We are interested in finding efficient estimates for the expression $\mathbb{E}\{(\Delta_n^t Y_n)(G_n^t \Delta_n)\}$. Such expressions are often encountered in signal processing applications, in particular, if $Y_n = G_n = X_n$ then this leads to the well known Excess Mean Square Error. We have now the following theorem:

Theorem 4

Let assumptions $\mathscr{A}1$, $\mathscr{A}2'$ and $\mathscr{A}3$ be valid and the matrix $\mathbf{A} = \mathbb{E}\{Z_1 X_1^t\}$ have eigenvalues in the left complex half plane, if the vector functions $Y(\xi)$, $G(\xi)$ satisfy

$$\sup_{\xi} \frac{\| Y(\xi) \| \, \| G(\xi) \|}{V(\xi)} < \infty \tag{35}$$

then

$$| \mathbb{E}\{(\Delta_n^t Y_n)(G_n^t \Delta_n)\} - \operatorname{trace}\{\Pi_n \mathbb{E}\{Y_1 G_1^t\}\} | = o(1) \left\{ \mu + \frac{\gamma^2}{\mu} + (1 - \mu\alpha)^n \right\} \tag{36}$$

where $\lim_{\mu \to 0} o(1) = 0$, uniformly in n.

Proof. The proof is based on Corollary 1, its main steps are presented in Appendix B. □

If we select $Y(\xi)$, $G(\xi)$ to be all possible combinations of constant vectors that have all their elements equal to zero except one that is unity then condition (35) is trivially satisfied and we have the following corollary that establishes $\mathbf{\Pi}_n$ as an efficient estimate of the covariance matrix $\mathbb{E}\{\Delta_n \Delta_n^t\}$.

Corollary 3

Under the assumptions of Theorem 4 we have that

$$\| \mathbb{E}(\Delta_n \Delta_n^t\} - \mathbf{\Pi}_n \| = o(1) \left\{ \mu + \frac{\gamma^2}{\mu} + (1 - \mu\alpha)^n \right\} \tag{37}$$

This result corresponds for example to Theorem 3.7 of Reference 13 (for the stationary case). What is also interesting to note by combining Theorem 4 and Corollary 3 is that, to a first order approximation, we can separate the expectation of Δ_n from quantities that depend only on the information supplied at time n. For example we can write $\mathbb{E}\{(\Delta_n \Delta_n^t)(Y_n G_n^t)\} \approx \mathbb{E}\{\Delta_n \Delta_n^t\} \mathbb{E}\{Y_n G_n^t\}$. This property was repeatedly used in the past, for the computation of second order statistics in adaptive algorithms, by invoking the Independence Assumption (Reference 20, p. 396).

4.3. The LMS algorithm

Let us apply the theory developed in the previous subsections to LMS so as to compare our results with Reference 15. For LMS we know that $Z_n = X_n$ and, according to our modelling, we also assume that $X_n = X(\xi_n)$ where $X(\xi)$ is some vector transformation of the underlying Markov proces $\{\xi_n\}$. Let us also for simplicity assume that $\{\xi_n\}$ is a Gaussian process generated by the state space model described in Section 3.1. As we have seen, we can then select as drift function the function $V(\xi) = \mathrm{e}^{\delta \xi^t \mathbf{S} \xi}$ for some $\delta > 0$.

In order to apply our results we need to satisfy Assumption $\mathscr{A}2'$ which here takes the form

$$\mathbb{E}_{\xi}\{ \| X(\xi_1) \|^{2l} V(\xi_1)\} \leqslant cV(\xi), \quad l = 1, 2 \tag{38}$$

This condition guarantees validity for the approximation of the convergence rate in (33) and for the approximation of the covariance matrix of Δ_n in (37). If we also like to apply Theorem 4 to estimate the excess mean square error we must impose the additional constant

$$\sup_{\xi} \frac{\| X(\xi) \|^2}{V(\xi)} < \infty \tag{39}$$

that corresponds to (35) with $Y_n = G_n = X_n$.

As we have seen in Section 3.1, there exists a $\delta' > 0$ such that if

$$\sup_{\xi} \frac{\| X(\xi) \|^{2l}}{e^{\delta' \xi^t \mathbf{S} \xi}} < \infty, \quad l = 1, 2 \tag{40}$$

then (38) is satisfied. Notice now that if $X(\xi)$ is either a polynomial vector transformation (or a vector transformation that can be bounded by polynomial) or even a vector transformation that grows exponentially in $\| \xi \|$ then both conditions (39) and (40) are satisfied. In other words if our data X_n are ANY polynomial or even exponential transformation of a Gaussian Markov process of the form of (22) then all our results can be applied. This should be compared with Reference 15 where the conditions imposed guarantee validity of (37) for Gaussian data whereas if X_n is any polynomial transformations of Gaussian random variables then the corresponding conditions fail and the results are not applicable.

5. CONCLUSION

By properly combining recent results concerning geometric ergodicity of Markov processes and long existing results concerning perturbations of linear operators we were able to propose a novel methodology for studying products of random matrices and in particular obtain efficient estimates for their performance. These estimates were consequently applied to the study of constant step size adaptive algorithms in order to establish estimates for their first and second order statistics. The contribution of our approach consists in enlarging considerably the class of combinations of data types and algorithms that can be studied by significantly relaxing certain restrictive conditions imposed on the data by existing methodologies. Furthermore, with our analysis we were also able to propose efficient estimates for the convergence rate of adaptive algorithms during the transient phase as compared to existing methods (not relying on the Independence Assumption) that can provide only bounds. Finally, it should be noted that we focused our analysis to data that are non-linear transformations of stationary Markov processes only for simplicity. Future publications are expected to generalize these results to non-stationary mixing processes as well.

APPENDIX A

Before presenting the proofs for Theorems 1 and 2 we require several definitions and preliminary results. We must stress that we are going to heavily rely on the theory existing in Reference 19 regarding perturbations of linear operators.

In order to prove our theorems we must first characterize the eigenstructure of the linear operator $\mathscr{T}(\mu)$ defined in (9). When we consider infinite dimensional linear operators the best way

to do this is through the use of the *resolvant operator* (Reference 19, p. 173). For z a complex number, the resolvant of a linear operator $\mathscr{M}$ is defined as $\mathscr{R}_{\mathscr{M}}(z) = (z1 - \mathscr{M})^{-1}$, that is, the inverse mapping of $z1 - \mathscr{M}$ where 1 stands for the identity operator. The set of complex points z where the resolvant $\mathscr{R}_{\mathscr{M}}(z)$ is bounded (in the $\|\cdot\|_V$ norm) is called the *resolvant set* whereas its complement is called the *spectrum* of the operator (Reference 19, pp. 173–174). The eigenvalues (if they exist) belong to the spectrum, the opposite is not necessarily true, that is, not every point in the spectrum is an eigenvalue (however this statement is true in the finite dimensional case Reference 19, p. 38).

We will now summarize in the form of lemmas two basic results from linear operator theory that we are going to need for our analysis.

Lemma 1

Let $\mathscr{M}$ be a bounded linear operator mapping $\mathscr{L}_V^\infty$ to $\mathscr{L}_V^\infty$ and $\Sigma(\mathscr{M})$ denote its spectrum then

$$\lim_{n\to\infty} (\|\mathscr{M}^n\|_V)^{1/n} = \sup_{z\in\Sigma(\mathscr{M})} |z| \tag{41}$$

Proof. The proof can be found in Reference 19, p. 176. □

The limit in (41) is known as the *spectral radius* of the operator (Reference 19, p. 176).

Lemma 2

Let the bounded linear operator $\mathscr{M}$ have spectrum $\Sigma(\mathscr{M})$ that can be written as the union of two parts $\Sigma(\mathscr{M}) = \Sigma_{\mathscr{F}}(\mathscr{M}) \cup \Sigma_{\mathscr{H}}(\mathscr{M})$ and assume that $\Sigma_{\mathscr{H}}(\mathscr{M})$ can be separated from $\Sigma_{\mathscr{F}}(\mathscr{M})$ by a simple closed curve. We can then define two operators $\mathscr{F}$, $\mathscr{H}$ with spectrum $\Sigma_{\mathscr{F}}(\mathscr{M})$, $\Sigma_{\mathscr{H}}(\mathscr{M})$, respectively, satisfying

$$\mathscr{M}^n = \mathscr{F}^n + \mathscr{H}^n \tag{42}$$

Proof. By combining Reference 19, pp. 172, 178 we have that we can find a projection $\mathscr{Q}$ (i.e. $\mathscr{Q}^2 = \mathscr{Q}$) such that $\mathscr{F} = \mathscr{M}\mathscr{Q}$ and $\mathscr{H} = \mathscr{M}(1 - \mathscr{Q})$. Furthermore, $\mathscr{Q}$ commutes with $\mathscr{M}$, that is, $\mathscr{Q}\mathscr{M} = \mathscr{M}\mathscr{Q}$ resulting in

$$\mathscr{Q}\mathscr{M}(1 - \mathscr{Q}) = (1 - \mathscr{Q})\mathscr{M}\mathscr{Q} = 0 \tag{43}$$

which yields $\mathscr{F}\mathscr{H} = \mathscr{H}\mathscr{F} = 0$. Consequently we can easily show (42). □

We can now proceed with the proof of Theorem 1.

Proof of Theorem 1. A key point in analysing the behaviour of $\mathscr{T}(\mu)$, defined in (12), is to observe that this linear operator can be regarded as a perturbed version of $\mathscr{T} = \mathscr{T}(0) = \mathbb{E}_\xi\{\cdot\}$. Consequently, we can relate the eigenstructure of $\mathscr{T}(\mu)$ to the corresponding of $\mathscr{T}$. In order to do so we first need to examine the eigenstructure of $\mathscr{T}$. Consider the resolvant $\mathscr{R}(z) = (z1 - \mathscr{T})^{-1}$ of $\mathscr{T}$. If $\mathscr{P} = \mathbb{E}\{\cdot\}$ denotes the projection defined by the steady state expectation then $\mathscr{T}\mathscr{P} = \mathscr{P}\mathscr{T} = \mathscr{P}$ and we can write

$$\mathscr{R}(z) = \frac{\mathscr{P}}{z - 1} + \bar{\mathscr{R}}(z) \tag{44}$$

where $\bar{\mathscr{R}}(z) = (z1 - \mathscr{T})^{-1}(1 - \mathscr{P})$. That $\mathscr{R}(z)$ can indeed be defined this way can be easily seen by directly verifying that $(z1 - \mathscr{T})\mathscr{R}(z) = 1$. Notice now that $\bar{\mathscr{R}}(z)$ exists (is bounded in the $\|\cdot\|_V$ norm) for at least all z satisfying $|z| > \rho$. This is so because $\bar{\mathscr{R}}(z)$ can be computed through the series $(z^{-1}1 + z^{-2}\mathscr{T} + z^{-3}\mathscr{T}^2 + \cdots)(1 - \mathscr{P})$ which, because of Assumption $\mathscr{A}1$, is convergent for all $|z| > \rho$ in the norm $\|\cdot\|_V$. Consequently, we observe from (44) that $z = 1$ is an isolated point of the spectrum $\Sigma(\mathscr{T})$ whereas all other points of this spectrum are inside the circle with radius ρ. In other words, the spectrum of $\mathscr{T}$ is comprised of two separate parts and thus $\mathscr{T}$, according to Lemma 2, can be decomposed into two operators. It is easy to see that the role of the projection $\mathscr{Q}$ plays here the projection $\mathscr{P}$ and we have that $\mathscr{F} = \mathscr{T}\mathscr{P} = \mathscr{P}$ and $\mathscr{H} = \mathscr{T}(1 - \mathscr{P})$ with the property $\mathscr{T}^n = \mathscr{F}^n + \mathscr{H}^n = \mathscr{P} + \mathscr{H}^n$. Furthermore, part $\mathscr{F} = \mathscr{P}$ has spectrum the single point $z = 1$ whereas $\mathscr{H}$ has its spectrum inside the circle with radius ρ. Because of (41) the term $\mathscr{H}^n$ tends in the $\|\cdot\|_V$ norm exponentially fast to zero at least as ρ^n.

Since $\mathscr{T}$ maps constant vectors of length K to themselves we conclude that $z = 1$ is an eigenvalue for $\mathscr{T}$ (this can also be deduced from the fact that $z = 1$ is an isolated point of the spectrum (Reference 19, p. 181). In other words the space $\mathbb{R}^K$ of constant vectors of constant vectors of length K, is an eigenspace of $\mathscr{T}$ with corresponding eigenvalue equal to unity and with multiplicity K. It is in fact $\mathscr{P} = \mathbb{E}\{\cdot\}$ the projection that projects into this finite dimensional eigenspace and therefore $\mathscr{P}$ is also an eigenprojection since it corresponds to the unit eigenvalue (Reference 19, p. 41).

As was stated above, our goal is to relate the eigenstructure of $\mathscr{T}(\mu)$ to the corresponding of $\mathscr{T} = \mathscr{T}(0) = \mathbb{E}_\xi\{\cdot\}$. Such a relation is particularly easy to establish whenever $\mathscr{T}(\mu)$ constitutes a holomorphic family (in μ) of operators. This important property is assured through Assumption $\mathscr{A}2$ as we show with the next lemma.

Lemma 3

Let the matrix functions $\mathbf{T}_l(\xi), l = 1, \ldots, r$, entering in the definition of the operators $\mathscr{T}^{(l)}$ in (11) satisfy Assumption $\mathscr{A}2$ then, for bounded μ, the linear operator $\mathscr{T}(\mu) = \mathscr{T} + \mu\mathscr{T}^{(1)} + \cdots + \mu^r\mathscr{T}^{(r)}$ is bounded and holomorphic of type (A) in μ.

Proof. Because of $\mathscr{A}2$ it is easy to prove that $\mathscr{T}(\mu)$ is bounded. According to Reference 19, p. 375, a family $\mathscr{T}(\mu)$ of operators is holomorphic of type (A) if (i) the domain of definition of $\mathscr{T}(\mu)$ is independent of μ and (ii) the vector $\mathscr{T}(\mu)G(\xi)$ as a function of μ is a holomorphic vector for every $G(\xi)$ in the domain of definition.

Condition (i) is true since the domain of $\mathscr{T}(\mu)$ is always $\mathscr{L}_V^\infty$. Furthermore Condition (ii) is also valid because $\mathscr{T}(\mu)G(\xi)$ is a polynomial vector function in μ with bounded (because of $\mathscr{A}2$) terms which is differentiable in μ, therefore from Reference 19, p. 365, we have that it is holomorphic. □

Continuing now with the proof of Theorem 1 let us consider, in the complex plane, the circle $|z| = \rho$. As we have seen this circle separates the spectrum of $\mathscr{T}$. Since $\mathscr{T}(\mu)$, according to Lemma 3, is holomorphic of type (A) we have from Reference 19, p. 379, Remark 2.9, that, for small enough μ, the spectrum of $\mathscr{T}(\mu)$ is also separated by the same circle. The decomposition of the spectrum suggests a corresponding decomposition of $\mathscr{T}(\mu)$ into two operators $\mathscr{T}(\mu) = \mathscr{F}(\mu) + \mathscr{H}(\mu)$ with $\mathscr{H}(\mu)$ corresponding to the part of the spectrum inside the circle and $\mathscr{F}(\mu)$ to the part of the spectrum outside the circle. This decomposition can be made possible with the help of a projection $\mathscr{P}(\mu)$ that commutes with $\mathscr{T}(\mu)$ and satisfies $\mathscr{F}(\mu) = \mathscr{T}(\mu)\mathscr{P}(\mu)$ and

$\mathscr{H}(\mu) = \mathscr{T}(\mu)[1 - \mathscr{P}(\mu)]$. Furthermore, because of (42), we have $\mathscr{T}(\mu)^n = \mathscr{F}(\mu)^n + \mathscr{H}(\mu)^n$ and because of (41) we can find c and $\bar{\rho}$ with $\rho < \bar{\rho} < 1$ such that

$$\| \mathscr{H}(\mu)^n \|_V \leqslant c\bar{\rho}^n \tag{45}$$

meaning that $\mathscr{H}(\mu)^n$ tends exponentially fast to zero at a rate that is independent of μ (for small enough μ).

What is interesting to note is the fact that although $\mathscr{P}$ is an eigenprojection for $\mathscr{T}$ (since it corresponds to its isolated unit eigenvalue) this is not necessarily the case for the operator $\mathscr{P}(\mu)$ which is the perturbed version of $\mathscr{P}$. This is so because, as we will soon see, the unit eigenvalue can lead, under perturbation, to more than one distinct eigenvalue and therefore the projection $\mathscr{P}(\mu)$ becomes a combination of eigenprojections.

From Reference 19, p. 212, we have that the subspace defined by the operator $\mathscr{F}(\mu)$ (or the projection $\mathscr{P}(\mu)$) has the same dimension as $\mathscr{P}$ and is thus finite dimensional with dimension equal to K. Consequently, its spectrum is comprised only of eigenvalues that are perturbed versions of the isolated unit eigenvalue of $\mathscr{T}$. To find estimates for the eigenvalues, eigenprojections and eigennilpotents of $\mathscr{F}(\mu)$, according to Reference 19, p. 379, Remark 2.10, we can apply results developed for the finite dimensional case.

Notice that we are interested in a group of eigenvalues of $\mathscr{T}(\mu)$ coming from perturbations of the isolated unit eigenvalue of $\mathscr{T}$ (a *λ-group*). Furthermore, for the unit eigenvalue of $\mathscr{T}$ we have that the corresponding eigennilpotent $\mathscr{D} = (\mathscr{T} - 1)\mathscr{P} = 0$, consequently this eigenvalue is *semisimple Reference* 19, p. 41. From Reference 19, p. 81, we then have the following approximation for the eigenvalues $f_i(\mu)$ of $\mathscr{F}(\mu)$:

$$f_i(\mu) = 1 + \mu\lambda_i + \mu o(1),\ i = 1, \ldots, s \tag{46}$$

where λ_i, $i = 1, \ldots, s$, are the eigenvalues of the finite dimensional operator $\mathscr{P}\mathscr{T}^{(1)}\mathscr{P}$ with $\mathscr{T}^{(1)}$ defined in (11). Recalling that $\mathscr{P} = \mathbb{E}\{\cdot\}$ and $\mathscr{T}^{(1)} = \mathbb{E}_\xi\{\mathbf{T}_1(\xi_1)\cdot\}$ it is easy to see that for $\mathbf{T} = \mathbb{E}\{\mathbf{T}_1(\xi_1)\}$ we have $\mathscr{P}\mathscr{T}^{(1)}\mathscr{P} = \mathbf{T}\mathscr{P} = \mathscr{P}\mathbf{T}$ (by verifying that all operators produce the same result on any element from $\mathscr{L}_V^\infty$). Therefore the eigenvalues of $\mathscr{P}\mathscr{T}^{(1)}\mathscr{P}$ coincide with the eigenvalues of the matrix $\mathbf{T}$ and this establishes (14).

Relation (15) is proved in Reference 19, p. 83, equation (2.45). To show (16) notice first that from Reference 19, p. 39, equation (5.21) we have that the eigenprojections $\mathscr{P}_i(\mu)$, $i = 1, \ldots, s$, satisfy $\mathscr{P}_i(\mu)\mathscr{P}_l(\mu) = \delta_{i,l}\mathscr{P}_i(\mu)$ where $\delta_{i,l}$ is the Kronecker delta and furthermore they decompose the total projection $\mathscr{P}(\mu)$ in the sense that $\mathscr{P}(\mu) = \mathscr{P}_1(\mu) + \cdots + \mathscr{P}_s(\mu)$. Since all $\mathscr{P}_i(\mu)$ commute with $\mathscr{T}(\mu)$ (because they are also eigenprojections of $\mathscr{T}(\mu)$) we can conclude that

$$[1 - \mathscr{P}(\mu)]\mathscr{T}(\mu)\mathscr{P}_i(\mu) = P_i(\mu)\mathscr{T}(\mu)[1 - \mathscr{P}(\mu)] = 0 \tag{47}$$

Using now induction in k for (16), relations (15), (46), (47), the fact that for any constant matrix $\mathbf{B}$ we can write $\mathbf{B}\mathscr{P} = \mathscr{P}\mathbf{B}$, after some straightforward computations and careful housekeeping we can show relation (16). □

Proof of Theorem 2. From Theorem 1 we have $\mathscr{T}(\mu)^n = \mathscr{F}(\mu)^n + \mathscr{H}(\mu)^n$ where $\| \mathscr{H}(\mu)^n \|_V \leqslant c\bar{\rho}^n$. This inequality can be also expressed as

$$\mathscr{H}(\mu)^n = \bar{\rho}^n O_V(1) \tag{48}$$

To show (18) we must prove that $\mathscr{F}(\mu)^n$ can be approximated by the remaining two terms in (18). The proof is based on the fact that the matrix $\mathbf{F}(\mu)^n$ can be decomposed in an exactly comparable

manner as the operator $\mathscr{F}(\mu)^n$ in (17) with eigenvalues, eigenvectors and eigennilpotents satisfying similar approximations as in Theorem 1 and show that $\mathscr{F}(\mu)^n - \mathbf{F}(\mu)^n\mathscr{P} = (1-\mu\alpha)^n o_V(1)$. This last relation can be proved by showing that a similar approximation applies to all terms entering in the decomposition of $\mathscr{F}(\mu)^n$. It should be noted however that this is possible only because we made the assumption that all eigenvalues have negative real part.

To show (19) it is sufficient to show that $\mathscr{P}\mathscr{H}(\mu)^n = \mu\bar{\rho}\mathscr{P}O_V(1)$. Because of (43) we have $\mathscr{P}(\mu)\mathscr{H}(\mu) = 0$ which yields $\mathscr{P}(\mu)\mathscr{H}(\mu)^n = 0$. Since, as we said above, $\mathscr{P}(\mu) = \mathscr{P}_1(\mu) + \cdots + \mathscr{P}_s(\mu)$ we conclude, using (14) and the fact that $\mathbf{P}_1 + \cdots + \mathbf{P}_s = \mathbf{I}_K$, that $\mathscr{P}(\mu) = \mathscr{P} + \mu O_V(1)$. Using also (48) we can write

$$\mathscr{P}\mathscr{H}(\mu)^n = \mu O_V(1)\mathscr{H}(\mu)^n = \mu\bar{\rho}^n O_V(1) \tag{49}$$

Multiplying this last relation from the left by $\mathscr{P}$ and using the property $\mathscr{P}^2 = \mathscr{P}$ yields the desired result. □

APPENDIX B

Proof of Theorem 3. We are going to apply Theorem 1 in order to show our result. Notice first that we can write

$$\mathbb{E}\{\Delta_n^t\} = \Delta_0^t\mathbb{E}\left\{\mathbb{E}_\xi\left\{\prod_{j=1}^{n}(\mathbf{I}_N - \mu X_j Z_j^t)\right\}\right\} = \Delta_0^t\mathscr{P}\mathscr{T}(\mu)^n\mathbf{I}_N \tag{50}$$

where $\mathscr{T}(\mu) = \mathbb{E}_\xi(\mathbf{I}_N - \mu X_1 Z_n^t\}$ (and the linear mapping of a matrix function is the collection of mappings of its columns). We can easily see that $\mathscr{A}2'$ implies the validity of $\mathscr{A}2$, therefore we can apply Theorem 1.

According to Theorem 1 the operator $\mathscr{T}(\mu)^n$ can be decomposed as $\mathscr{T}(\mu)^n = \mathscr{F}(\mu)^n + \mathscr{H}(\mu)^n$. Part $\mathscr{F}(\mu)^n$ is the leading one as compared to $\mathscr{H}(\mu)^n$ and we have that it can be further decomposed as in (17). Let us for simplicity assume that λ_1 is the eigenvalue of $\mathbf{A} = \mathbb{E}\{X_1 Z_1^t\}$ with the smallest real part and $\mathbf{P}_1$ the corresponding eigenprojection. Then $f_1(\mu)$, because of (14), will have for small enough μ the maximum amplitude among all eigenvalues of $\mathscr{F}(\mu)$. Moreover if we assume that $\Delta_0^t\mathbf{P}_1 \neq 0$ (i.e. the initial condition excites the eigenvalue with the maximum amplitude) we conclude that

$$\mathbb{E}\{\Delta_n^t\} = \Delta_n^t\mathscr{P}\mathscr{T}(\mu)^n\mathbf{I}_N = f_1(\mu)^n n^r\mu^r\Theta^t(1) \tag{51}$$

for some r with $0 \leqslant r \leqslant m_1$ and m_1 the multiplicity of $f_1(\mu)$ and $\Theta(1)$ a vector that has a norm which, for small enough μ, is uniformly away from zero and infinity. Taking logarithm and the limit after dividing by n we obtain

$$\lim_{n\to\infty}\frac{-\log(\|\mathbb{E}(\Delta_n\}\|)}{n} = -\log(|f_1(\mu)|) = -\log(|1-\mu\lambda_1+\mu o(1)|) = \mu\mathrm{Re}(\lambda_1) + \mu o(1) \tag{52}$$

And this concludes the proof □

Proof of Theorem 4. To prove the theorem notice first that we can write

$$(\Delta_n^t Y_n)(G_n^t\Delta_n) = (\Delta_n^t \otimes \Delta_n^t)(Y_n \otimes G_n) \tag{53}$$

where ' $\otimes$ ' denotes Kronecker product. Using (31) and stationarity we conclude

$$\mathbb{E}\{(\Delta_n^t \otimes \Delta_n^t)(Y_n \otimes G_n)\} = (\Delta_n^t \otimes \Delta_n^t)\mathbb{E}\left\{\left[\prod_{j=1}^{n}(\mathbf{I}_N - \mu X_j Z_n^t) \otimes (\mathbf{I}_N - \mu X_j Z_j^t)\right](Y_n \otimes G_n)\right\}$$
$$+ \mathbb{E}\left\{\sum_{j=1}^{n}(\mu^2\sigma_w^2 Z_0^t \otimes Z_0^t + \gamma^2 \operatorname{vect}\{\mathbf{Q}_U)^t \times \left[\prod_{l=1}^{n-j}(\mathbf{I}_N - \mu X_l Z_l^t) \otimes (\mathbf{I}_N - \mu X_l Z_l^t)\right](Y_{n-j} \otimes G_{n-j})\right\} \tag{54}$$

We observe that we can again apply our results to (54) by defining the operator $\mathscr{T}(\mu) = \mathbb{E}_\xi\{\mathbf{I}_N - \mu X_1 Z_1^t) \otimes (\mathbf{I}_N - \mu X_1 Z_1^t)$.$\}$ with dimension K being equal to $K = N^2$. Again with the help of condition $\mathscr{A}2'$ we can show that $\mathscr{A}2$ is satisfied, moreover (35) insures that $Y(\xi) \otimes G(\xi)$ belongs to $\mathscr{L}_V^\infty$, consequently we can use Corollary 1. For the approximation of $\mathscr{T}(\mu)^n$ we can use the matrix $\mathbf{F}(\mu) = (\mathbf{I}_N - \mu\mathbf{A}^t) \otimes (\mathbf{I}_N - \mu\mathbf{A}^t)$ since it has the form suggested by the corollary. More specifically we need to use (19) for the first term in the rhs of (54) and (18) for the second term.

On the other hand we can write

$$\operatorname{trace}\{\mathbf{\Pi}_n \mathbb{E}\{Y_1 G_1^t\}\} = \operatorname{vect}\{\mathbf{\Pi}_n\}^t \mathbb{E}\{Y_1 \otimes G_1\} \tag{55}$$

Using (34), we obtain

$$\operatorname{vect}\{\mathbf{\Pi}_n\}^t = \operatorname{vect}\{\mathbf{\Pi}_{n-1}\}^t(\mathbf{I}_N - \mu\mathbf{A}^t) \otimes (\mathbf{I}_N - \mu\mathbf{A}^t) + \mu^2\sigma_w^2 \operatorname{vect}\{\mathbf{Q}_Z\}^t + \gamma^2 \operatorname{vect}\{\mathbf{Q}_U\}^t \tag{56}$$

meaning that

$$\operatorname{vect}\{\mathbf{\Pi}_n\}^t = (\Delta_0^t \otimes \Delta_0^t)\mathbf{F}(\mu)^n + \sum_{j=1}^{n}(\mu^2\sigma_w^2 \operatorname{vect}\{\mathbf{Q}_Z\} + \gamma^2 \operatorname{vect}\{\mathbf{Q}_U\})^t \mathbf{F}(\mu)^{n-j} \tag{57}$$

with the matrix $\mathbf{F}(\mu)$ defined above. Applying now Corollary 1 to (54) in the way we described earlier we can compare the outcome with (55) after using (57). It is then straightforward to see that relation (36) is indeed true. □

REFERENCES

1. Ljung, L. 'Analysis of recursive stochastic algorithms', *IEEE Trans. Autom. Control* **22**(4) 551–575 (1977).
2. Benveniste, A., M. Metivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer, New York, 1990.
3. Bucklew, J.A., T.G. Kurtz and W.A. Sethares, 'Weak convergence and local stability properties of fixed step size recursive algorithms *IEEE Trans. Inform. Theory*, **IT-39** (3) 966–978 (1993).
4. Eweda, E. and O. Macchi, 'Tracking error bounds of adaptive nonstationary filtering', *Automatica*, **21**(3) 293–302 (1985).
5. Kushner, H.J., *Approximation and Weak Convergence Methods for Random Processes*, MIT Press Series in Signal Processing, Optimization and Control, MIT Press, Cambridge, MA, 1984.
6. Kushner, H.J. and D.S. Clark, *Stochastic Approximation Methods for Constraint and Unconstraint Systems*, Springer, New York, 1978.
7. Ljung, L. and S. Gunnarsson, 'Adaptive tracking in system identification, a survey', *Automatica*, **26**(1), 7–22 (1990).
8. Ljung, L. and P. Priouret, 'A result on the mean square tracking error', *Int. J. Adapt. Control Signal Process.* **5**(4), 231–250 (1991).
9. Ljung, L. and P. Priouret, 'Remarks on the mean square tracking error', *Int. J. Adapt. Control Signal Process.* **5**(6), 395–403 (1991).

10. Solo, V. and X. Kong, *Adaptive Signal Processing Algorithms, Stability and Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
11. Guo, L. 'Stability of recursive tracking algorithms', SIAM J. *Control Optim.*, **32**(5), 1195–1225 (1994).
12. Guo, L. and L. Ljung, 'Exponential stability of general tracking algorithms', *IEEE Trans. Autom. Control* **40**(8), 1376–1387 (1995).
13. Gou, L. and L. Ljung, 'Performance analysis of general tracking algorithms', *IEEE Trans. Autom. Control*, **40**(8), 1388–1402 (1995).
14. Guo, L., L. Ljung and P. Priouret, 'Performance analysis of forgetting factor RLS', *Int. J. Adapt. Control Signal Process.*, **7,** 525–537 (1993).
15. Guo, L., L. Ljung and G.L. Wang, 'Necessary and sufficient conditions for stability of LMS', *IEEE Trans. Autom. Control*, **42**(6), 761–770.
16. Juditski, A. and P. Priouret, 'A robust algorithm for random parameter tracking', *IEEE Trans. Autom. Control* **39**(6), 1211–1221 (1994).
17. Niedźwiecki, M. and L. Guo, 'Nonasymptotic results for finite memory WLS filters', *IEEE Trans. Autom. Control*, **36**(2) 198–206 (1991).
18. Meyn, S.P. and R.L. Tweedie, *Markov Chains and Stochastic Stability*, Springer, New York, 1993.
19. Kato, T., *Perturbation Theory for Linear Operators*, Springer, New York, 1966.
20. Haykin, S., *Adaptive Filter Theory*, 3-rd edn, Prentice-Hall, Englewood Cliffs, NJ, 1996.

