RECURSIVE IDENTIFICATION AND ADAPTIVE PREDICTION
IN LINEAR STOCHASTIC SYSTEMS

BY

TZE LEUNG LAI AND ZHILIANG YING

TECHNICAL REPORT NO. 12
NOVEMBER 1989

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS87-15614

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
RECURSIVE IDENTIFICATION AND ADAPTIVE PREDICTION
IN LINEAR STOCHASTIC SYSTEMS

BY

TZE LEUNG LAI AND ZHILIANG YING

TECHNICAL REPORT NO. 12
NOVEMBER 1989

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS87-15614

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
RECURSIVE IDENTIFICATION AND ADAPTIVE PREDICTION IN LINEAR STOCHASTIC SYSTEMS*

TZE LEUNG LAI† AND ZHILIANG YING‡

*This work was supported by the National Science Foundation, the National Security Agency and the Air Force Office of Scientific Research.
† Department of Statistics, Stanford University, Stanford, California 94305.
‡ Department of Statistics, University of Illinois, Champaign, Illinois 61820.
Abstract. By making use of extended stochastic Lyapunov functions and martingale limit theorems, we establish herein certain basic properties of adaptive $d$-step ahead predictors associated with the extended least squares, stochastic gradient (without interlacing), and monitored recursive maximum likelihood algorithms for recursive identification of an ARMAX system. Both the direct (or implicit) and indirect (or explicit) approaches to adaptive prediction are considered within a unified framework involving stochastic regression models. Applications to adaptive control of ARMAX systems are also discussed.

Key words: Adaptive prediction, global convergence, stochastic gradient algorithm, AML, recursive MLE, stochastic adaptive control, certainty equivalence, asymptotic efficiency.

AMS(MOS) subject classifications. 93E12, 93C40, 60G42.
1. Introduction and background. Consider the ARMAX system (autoregressive moving average system with exogenous inputs) defined by the linear stochastic difference equation

\begin{equation}
A(q^{-1})y_n = q^{-\Delta}B(q^{-1})u_n + C(q^{-1})\epsilon_n,
\end{equation}

where \( \{y_n\} \), \( \{u_n\} \) and \( \{\epsilon_n\} \) denote the output, input and disturbance sequences, respectively, \( \Delta \geq 1 \) represents the delay and

\begin{equation}
A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_p q^{-p}, \quad B(q^{-1}) = b_1 + \cdots + b_k q^{-(k-1)}, \quad C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_h q^{-h}
\end{equation}

are scalar polynomials in the backward shift operator \( q^{-1} \). Throughout the sequel we shall assume that the sequence \( \{\epsilon_n\} \) is a martingale difference sequence with respect to an increasing sequence of \( \sigma \)-fields \( \mathcal{F}_n \) such that

\begin{equation}
\sup_n E(|\epsilon_n|^\alpha |\mathcal{F}_{n-1}) < \infty \quad \text{a.s. for some } \alpha > 2.
\end{equation}

Moreover, the input \( u_t \) at stage \( t \) is assumed to be \( \mathcal{F}_t \)-measurable (i.e., involving only the current and past observations \( y_t, y_{t-1}, u_{t-1}, \cdots \), but no future observations). Letting \( x_0 = (y_0, \cdots, y_{-p}, u_0, \cdots, u_{2-\Delta-k}, \epsilon_0, \cdots, \epsilon_{-h}) \) denote the "initial condition" of (1.1), it is also assumed that \( x_0 \) is \( \mathcal{F}_0 \)-measurable.

Let \( 1 \leq d \leq \Delta \). When the system parameters \( a_1, \cdots, a_p, b_1, \cdots, b_k, c_1, \cdots, c_h \) and the initial condition \( x_0 \) are known, the minimum variance \( d \)-step ahead predictor \( \tilde{y}_{n+d} \triangleq E(y_{n+d} | \mathcal{F}_n) \) of the output \( y_{n+d} \) can be determined recursively by the Åström predictor identity (1.6) below (cf. [1], [2], [3], [4]). By the division algorithm, there exist polynomials \( F(z) = 1 + f_1 z + \cdots + f_{d-1} z^{d-1} \) and \( G(z) = g_1 + \cdots + g_{p(d)} z^{p(d)-1} \) with \( p(d) = p \lor (h - d + 1) \) such that

\begin{equation}
C(z) = F(z)A(z) + z^d G(z),
\end{equation}

and therefore (1.1) can be rewritten in the form

\begin{equation}
C(q^{-1})\{y_{n+d} - F(q^{-1})\epsilon_{n+d}\} = G(q^{-1})y_n + q^{-(\Delta-d)}F(q^{-1})B(q^{-1})u_n.
\end{equation}

This implies that the minimum variance predictor \( \tilde{y}_{n+d} \) is given recursively by

\begin{equation}
C(q^{-1})\tilde{y}_{n+d} = G(q^{-1})y_n + q^{-(\Delta-d)}F(q^{-1})B(q^{-1})u_n.
\end{equation}

The prediction error of the predictor \( \tilde{y}_{n+d} \) is

\begin{equation}
\eta_{n+d} \triangleq y_{n+d} - \tilde{y}_{n+d} = F(q^{-1})\epsilon_{n+d}.
\end{equation}
In practice, the system parameters and initial condition are usually unknown, and one has to "adapt" the optimal predictor (1.6) by substituting the unknown entities in (1.6) by their estimates. The so-called explicit (or indirect) approach of adaptive prediction is to first estimate the parameters \(a_1, \ldots, a_p, b_1, \ldots, b_k, c_1, \ldots, c_h\) of the explicit dynamical system (1.1) and then to substitute these parameter values that appear in the polynomials \(C(q^{-1}), B(q^{-1}), F(q^{-1})\) and \(G(q^{-1})\) of (1.6) by their estimates. In contrast, the implicit (or direct) approach of adaptive prediction is to first develop directly recursive estimates \(\theta_n\) of the parameter vector

\[
\theta = (g_1, \ldots, g_{p(d)}, b_1, (fb)_2, \ldots, (fb)_{k+d-1}, -c, \ldots, -c_h)', \quad \text{where}
\sum_{i=1}^{k+d-1} (fb)_i z^{i-1} = F(z)B(z), \quad \text{so that} \quad (fb)_1 = b_1,
\]

of the system's implicit representation that combines (1.5) and (1.7) into the form

\[
y_{n+d} = \theta'\psi_n + \eta_{n+d}, \quad \text{where} \quad \psi_n = (y_n, \ldots, y_{n-p(d)+1}, u_{n-\Delta+d}, \ldots, u_{n-k-\Delta+2}, \tilde{y}_{n+d-1}, \ldots, \tilde{y}_{n+d-h})'.
\]

Letting \(\hat{y}_{n+d}\) denote the adaptive predictor of \(y_{n+d}\), this implicit approach generates \(\hat{y}_{n+d}\) recursively by

\[
\hat{y}_{n+d} = \theta'_n \phi_n, \quad \text{where} \quad \phi_n = (y_n, \ldots, y_{n-p(d)+1}, u_{n-\Delta+d}, \ldots, u_{n-k-\Delta+2}, \tilde{y}_{n+d-1}, \ldots, \tilde{y}_{n+d-h})'.
\]

For the explicit approach, there is a large literature on recursive estimation of the parameter vector

\[
\Theta = (-a_1, \ldots, -a_p, b_1, \ldots, b_k, c_1, \ldots, c_h')
\]

of the dynamical system (1.1), which can be written in the regression form

\[
y_n = \Theta'\Psi_{n-1} + \epsilon_n, \quad \text{where} \quad \Psi_t = (y_t, \ldots, y_{t-p+1}, u_{t-\Delta+1}, \ldots, u_{t-\Delta-k+2}, \epsilon_t, \ldots, \epsilon_{t-h+1})'.
\]

The recent monographs [2], [3], [4] provide excellent unified overviews of various recursive estimation algorithms in the literature. In particular, an important and widely studied problem concerning these recursive estimators is under what conditions they converge. In the seminal papers [5], [6], Ljung developed the ODE (ordinary differential equation) method for the convergence analysis of recursive estimators \(\Theta_n\) of \(\Theta\). Under certain \textit{a priori} boundedness and recurrence assumptions on \(\Theta_n\), this method introduces a space-time renormalization into the recursion for \(\Theta_n - \Theta\) to obtain a nonrandom ODE as a limit point and studies the limiting
behavior of $\Theta_n$ via the stability properties of the associated ODE. Such stability analysis is often conveniently carried out by making use of a Lyapunov function. Instead of working with a Lyapunov function associated with the limiting ODE, an obvious alternative is to develop an analogue for the original recursions defining $\Theta_n$. This is the idea behind the “stochastic Lyapunov function” approach introduced by Moore and Ledwich [7] and Solo [8]. A basic ingredient of this approach is to use the underlying system dynamics to develop recursive inequalities for a suitably chosen nonnegative random function of $\Theta_n$ and to normalize and transform this function into a nonnegative almost supermartingale (stochastic Lyapunov function) to which the martingale convergence theorem can be applied. In particular, for the AML algorithm

\begin{align*}
(1.13a) \quad \Theta_n &= \Theta_{n-1} + P_{n-1} \Phi_{n-1}(y_n - \Theta'_{n-1} \Phi_{n-1}), \quad P_n^{-1} = P_{n-1}^{-1} + \Phi_n \Phi'_n, \\
(1.13b) \quad \Phi_n &= (y_n, \ldots, y_{n-p+1}, u_{n-\Delta+1}, \ldots, u_{n-\Delta-k+2}, \hat{e}_n, \ldots, \hat{e}_{n-h+1}), \\
(1.13c) \quad \hat{e}_n &= y_n - \Theta'_{n} \Phi_{n-1},
\end{align*}

Solo [8] used this approach to prove the strong consistency of $\Theta_n$ under both the “persistent excitation” condition

\begin{equation}
(1.14) \quad n^{-1} \sum_{i=1}^{n} \Psi_i \Psi'_i \text{ converges a.s. to a positive definite matrix,}
\end{equation}

and the “positive real” condition

\begin{equation}
(1.15) \quad C(z) \text{ is stable and } \min_{|z|=1} \text{Re}(1/C(z) - 1/2) > 0.
\end{equation}

Here and in the sequel we say that a polynomial $C(z)$ is “stable” if all its zeros lie outside the unit circle.

For the AML algorithm, by using martingale limit theorems (not restricted to convergence) to analyze directly Solo’s recursive inequalities for the quadratic form

\begin{equation}
(1.16) \quad Q_n \overset{\Delta}{=} (\Theta_n - \Theta)' P_{n-1}^{-1} (\Theta_n - \Theta),
\end{equation}

instead of following Solo [8] to transform $Q_n$ into a nonnegative almost supermartingale that converges a.s. by the martingale convergence theorem, Lai and Wei [9] established the strong consistency of $\Theta_n$ under (1.15) and the much weaker excitation condition

\begin{equation}
(1.17) \quad \lambda_{\min} \left( \sum_{i=1}^{n} \Psi_i \Psi'_i \right) \to \infty \quad \text{and} \quad \log \lambda_{\max} \left( \sum_{i=1}^{n} \Psi_i \Psi'_i \right) = o \left( \lambda_{\min} \left( \sum_{i=1}^{n} \Psi_i \Psi'_i \right) \right) \text{ a.s.}
\end{equation}

Here and in the sequel we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the maximum and minimum eigenvalues of a symmetric matrix $A$. Because of the Lyapunov-type recursive inequalities
satisfied by $Q_n$, which, however, need not be convergent, we shall call such functions “extended stochastic Lyapunov functions” as in [10], where it is shown that stronger results can often be obtained by applying martingale theory directly to such functions without transforming them into (convergent) stochastic Lyapunov functions.

Although the recursive estimator (1.13) has been called “approximate maximum likelihood” (AML), it does not arise from the maximization of the log-likelihood function, as in the off-line (non-recursive) maximum likelihood estimator, when the $\epsilon_i$ are assumed to be normally distributed with mean 0 and variance $\sigma^2$. The recursive maximum likelihood estimator RML2, introduced by Åström and Söderström, replaces (1.13a) by

\[(1.18a) \quad \Theta_n = \Theta_{n-1} + P_{n-1} \xi_{n-1} (y_n - \Theta'_{n-1} \Phi_{n-1}), \quad P_n^{-1} = P_{n-1}^{-1} + \xi_n \xi'_n,
\]

where letting $\Theta_n = (-\hat{a}_{n,1}, \ldots, -\hat{a}_{n,p}, \hat{b}_{n,1}, \ldots, \hat{b}_{n,k}, \hat{c}_{n,1}, \ldots, \hat{c}_{n,h})'$, define $\xi_n$ recursively by

\[(1.18b) \quad \xi_n + \hat{c}_{n-1,1} \xi_{n-1} + \cdots + \hat{c}_{n-1,h} \xi_{n-h} = \Phi_n,
\]

cf. [2]. This algorithm is based on first replacing the derivative of the log-likelihood function by its linear approximation around the true parameter $\Theta$ and then replacing the unknown $\Theta$ by $\Theta_{n-1}$, and would therefore lead to an asymptotically efficient estimator if $\Theta_{n-1}$ should converge to $\Theta$. In §5 we introduce an additional monitoring scheme to ensure that $\Theta_{n-1}$ is eventually close to $\Theta$, and use the extended Lyapunov function (1.16) to analyze this modification of the RML2 algorithm, which we call the “monitored recursive maximum likelihood algorithm”.

In the explicit approach to adaptive prediction, for a recursive algorithm $\Theta_n$ estimating the unknown parameter vector $\Theta$ defined in (1.11), we first use $\Theta_n$ at stage $n$ to estimate the unknown coefficients of the polynomials $C(q^{-1})$, $B(q^{-1})$, $F(q^{-1})$ and $G(q^{-1})$ in (1.6), leading to the estimated polynomials $\hat{C}_n(q^{-1})$, $\hat{B}_n(q^{-1})$, $\hat{F}_n(q^{-1})$ and $\hat{G}_n(q^{-1})$ at stage $n$, and then define the predictor $\hat{y}_{n+d}$ of $y_{n+d}$ by the recursive relation

\[(1.19) \quad \hat{C}_n(q^{-1}) \hat{y}_{n+d} = \hat{G}_n(q^{-1}) y_n + q^{-(\Delta - d)} \hat{F}_n(q^{-1}) \hat{B}_n(q^{-1}) u_n,
\]

noting that the coefficients of $F(q^{-1})$ and $G(q^{-1})$ are polynomial functions of the components of $\Theta$ by (1.4). Hence, if $\Theta_n$ converges a.s. to $\Theta$ and $C(z)$ is stable, then it follows from (1.6) and (1.19) that

\[(1.20) \quad \sum_{i=1}^{n} (\hat{y}_{i+d} - \hat{y}_{i+d})^2 = o\left(\sum_{i=1}^{n} (y_i^2 + u_i^2)\right) \text{ a.s.},
\]

cf. Lemma 5 of §2. In most applications, one typically has sample mean square boundedness for the input-output data, i.e.,

\[(1.21) \quad \limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} (y_i^2 + u_i^2) < \infty \text{ a.s.},
\]
in which case (1.20) implies that

\[(1.22) \quad n^{-1} \sum_{i=1}^{n} (\tilde{y}_{i+\Delta} - \hat{y}_{i+\Delta})^2 \to 0 \quad \text{a.s.}\]

A sequence of \(d\)-step ahead predictors \(\{\tilde{y}_{n+d}\}\) is said to be “globally convergent” if (1.22) holds (cf. [11]). We have pointed out above that if a consistent estimator \(\Theta_n\) of the parameter vector \(\Theta\) of the explicit system (1.11) can be found and if \(C(z)\) is stable and (1.21) holds, then globally convergent adaptive predictors can be constructed by the recursive relation (1.19). However, the requirement of consistency in parameter estimation is often not needed in the construction of globally convergent adaptive predictors, particularly if one uses an implicit approach. Extending the AML algorithm to the implicit model (1.9), Sin, Goodwin and Bitmead [12] constructed \(d\)-step ahead adaptive predictors based on an interlaced AML algorithm and showed that such predictors are globally convergent under the assumption (1.21) and an assumption analogous to (1.15). As pointed out by Zhang [13], however, their proof uses Solo’s [8] result (A6) whose proof contains a gap. In §4, where Solo’s result (A6) is shown to be incorrect without additional assumptions, we prove a stronger result than the Sin-Goodwin-Bitmead theorem under the additional assumption that \(\limsup_{n \to \infty} \phi_n' (\sum_{i=1}^{n} \phi_i \phi_i')^{-1} \phi_n < 1\) a.s., where the \(\phi_n\) are the pseudoregression vectors in their algorithm.

The global convergence property (1.22) for adaptive predictors is of particular interest in the adaptive control problem of setting the input \(u_t\) at stage \(t\) so that the output \(y_{t+\Delta}\) is as close as possible to some target value \(y^*_{t+\Delta}\). When the system parameters and the initial condition \(x_0\) are known and \(b_1 \neq 0\), the minimum variance controller is to set \(u_t\) such that \(\tilde{y}_{t+\Delta} = y^*_t + \Delta\). In ignorance of \(x_0\) and the system parameters, it is therefore natural to set \(u_t\) such that \(\tilde{y}_{t+\Delta} = y^*_t + \Delta\), where \(\tilde{y}_{t+\Delta}\) is a globally convergent adaptive \(\Delta\)-step ahead predictor of \(y_{t+\Delta}\). Since \(\tilde{y}_{t+\Delta} = y_{t+\Delta} - \eta_{t+\Delta}\), (1.22) implies the so-called “self-optimizing” property that

\[(1.23) \quad n^{-1} \sum_{i=\Delta+1}^{n} (y_i - y_i^* - \eta_i)^2 \to 0 \quad \text{a.s.,}\]

for the adaptive controller defined by \(\tilde{y}_{t+\Delta} = y^*_t + \Delta\).

In the case of unit delay \(\Delta = 1\), Goodwin, Ramadge and Caines [14] used this approach in conjunction with the stochastic gradient algorithm

\[(1.24) \quad \theta_n = \theta_{n-1} + (a/r_{n-1}) \phi_{n-1} (y_n - \tilde{y}_n), \quad r_n = r_{n-1} + \|\phi_n\|^2,\]

where \(\phi_t\) and \(\tilde{y}_t\) are defined in (1.10) (with \(d = 1\)) and \(a > 0\), to establish the self-optimizing property (1.23) for the adaptive controller that chooses the input \(u_t\) so that

\[(1.25) \quad \tilde{y}_{t+1} = y^*_{t+1},\]

\[5\]
under the assumptions

\[ C(z) \text{ is stable and } \min_{|z|=1} \text{Re}(1/C(z) - 1/2) > 0, \]

\[ B(z) \text{ is stable and } b_1 \neq 0, \]

\[ (x_0, \epsilon_1, \cdots, \epsilon_n) \text{ is absolutely continuous with respect to Lebesgue measure for every } n \geq 1. \]

The assumption (1.28) ensures that the component of \( \theta_n \) estimating the component \( b_1 \) of \( \theta \) is nonzero a.s. and therefore we can indeed define \( u_t \) by \( \theta'_t \phi_t = y_{t+1}^*, \) i.e., by (1.25). As shown in [11] for general delay \( \Delta \) but still for \( d = 1, \) (1.26) and (1.21) are sufficient conditions for the global convergence of adaptive 1-step ahead predictors based on the stochastic gradient algorithm (1.24). The assumption (1.27) is needed to ensure that (1.21) holds for the adaptive controller (1.25).

For general delay \( \Delta \) and \( d = \Delta, \) Goodwin, Sin and Saluja [15] extended (1.24) to the form

\[ \theta_n = \theta_{n-d} + (a/r_{n-d}) \phi_{n-d} (y_n - \theta'_{n-d} \phi_{n-d}), \quad r_n = r_{n-1} + \| \phi_n \|^2, \]

in which \( \phi_n \) is given by (1.10). Under the assumptions (1.26) and (1.21), they showed that the adaptive \( d \)-step head predictors \( \hat{y}_{n+d} = \theta'_n \phi_n \) are globally convergent. By using the explicit instead of the implicit approach, Fuchs [16], [17] also constructed globally convergent adaptive \( d \)-step ahead predictors of the form (1.19), in which the coefficients of the polynomials \( \hat{C}_n(q^{-1}), \hat{B}_n(q^{-1}), \hat{F}_n(q^{-1}) \) and \( \hat{G}_n(q^{-1}) \) are determined from the stochastic gradient algorithm \( \Theta_n \) estimating (1.11), defined by

\[ \begin{align*}
(1.30a) \quad \Theta_n &= \Theta_{n-1} + (a/r_{n-1}) \Phi_{n-1} \epsilon_n, \\
(1.30b) \quad \epsilon_n &= y_n - \Theta'_{n-1} \Phi_{n-1}, \\
(1.30c) \quad r_n &= r_{n-1} + \| \Phi_n \|^2, \\
& \quad \text{where} \quad \Phi_n = (y_n, \cdots, y_{n-p+1}, u_{n-\Delta+1}, \cdots, u_{n-\Delta-k+2}, \epsilon_n, \cdots, \epsilon_{n-h+1})'.
\end{align*} \]

Unlike the single recursion in (1.30), the algorithm (1.29) interlaces \( d \) recursions each of which is similar to a unit-delay algorithm. It has been an open problem concerning whether such multiple recursions are indeed necessary and not just dictated by the stochastic Lyapunov function method of convergence analysis, cf. [16, p. 219]. By using the extended stochastic Lyapunov function approach instead, we show in \S 3 that interlacing is not needed to establish global convergence of adaptive \( d \)-step ahead predictors based on the stochastic gradient algorithm for the implicit model (1.9), giving a positive answer to this long-standing open problem.

Although the stochastic gradient algorithm (with a scalar gain \( a/r_t \)) leads to globally convergent adaptive predictors by either the explicit or implicit approach under the assumptions
(1.21) and (1.26), the rate of convergence in (1.22) of such adaptive predictors is usually inferior to that associated with recursive identification algorithms using matrix gains, as noted by Lai, Wei and Zhang [18] who illustrated this point by the following simple example. Consider 1-step ahead prediction in the ARX system

\begin{equation}
    y_{n+1} = \alpha y_n + \beta u_n + \epsilon_{n+1},
\end{equation}

where \(|\alpha| < 1\), the \(\epsilon_n\) are independent normal random variables with mean 0 and variance \(\sigma^2 > 0\), and the inputs \(u_n\) are also independent normal random variables with mean 0 and variance \(\sigma_n^2 = n^{-2\gamma}\) for some \(0 < \gamma < 1/2\) and such that \(\{u_n\}\) and \(\{\epsilon_n\}\) are independent sequences. In this case, for the adaptive predictor \(\hat{y}_{n+1}^G = \Theta_n' \Phi_n\) defined by the stochastic gradient algorithm (1.30) with \(\alpha = 1\) and \(\Phi_n = (y_n, u_n)'\), the convergence rate in (1.22) cannot be faster than \(n^{-2\gamma}\) since

\begin{equation}
    P\left\{ \liminf_{n \to \infty} n^{-(1-2\gamma)} \sum_{i=1}^{n} (\tilde{y}_{i+1} - \hat{y}_{i+1}^G)^2 > 0 \right\} > 0.
\end{equation}

On the other hand, the adaptive predictor \(\hat{y}_{n+1}^{LS}\) associated with the least squares estimate \((\sum_1^{n-1} \Phi_i \Phi_i')^{-1} \sum_1^{n-1} \Phi_i y_{i+1}\) of \(\Theta\) satisfies

\begin{equation}
    \limsup_{n \to \infty} \sum_{i=1}^{n} (\tilde{y}_{i+1} - \hat{y}_{i+1}^{LS})^2 / \log n \leq \sigma^2 \quad \text{a.s.},
\end{equation}

which implies that the convergence rate in (1.22) is \(O(n^{-1} \log n)\), cf. [18, p.179].

In the ARX system (1.31) with i.i.d. normal errors \(\epsilon_n\), the least squares estimate coincides with the maximum likelihood estimate which is asymptotically efficient, and the logarithmic order in (1.33) is therefore asymptotically minimal for adaptive 1-step ahead predictors. To develop similar results for the \(d\)-step ahead prediction problem in ARMAX models with unknown parameters, we extend in §5 the RML2 algorithm (1.18) to the implicit system (1.9) and introduce an additional monitoring scheme to ensure that the recursive estimates \(\theta_n\) generated by the algorithm are eventually close to the parameter vector \(\theta\) of (1.9). By making use of extended stochastic Lyapunov functions, we are able to extend (1.33) to adaptive \(d\)-step ahead predictors based on \(\theta_n\).

In summary, the concept of extended stochastic Lyapunov functions provides a unified treatment of adaptive predictors based on various recursive identification algorithms, using either the explicit or the implicit approach. In particular, we use this idea in §3 to solve an open problem in the literature concerning the global convergence of adaptive \(d\)-step ahead predictors based on the stochastic gradient algorithm without interlacing in the implicit approach, to improve in §4 previous results on adaptive predictors based on extended least squares or modified least squares, and to obtain in §5 an analogue of (1.33) for the adaptive \(d\)-step ahead
prediction problem by using the monitored recursive maximum likelihood algorithm that generalizes the least squares algorithm in (1.33). These results, which will be extended to multivariable systems in §6, are of basic interest to adaptive control problems. As will be discussed in §6, while (1.22) leads to the self-optimizing property (1.23) of adaptive controllers based on the stochastic gradient algorithm, (1.33) and its extensions in §5 lead to the much stronger property \( \sum_{i=1}^{n} (y_i - y_i^* - \eta_i)^2 = O(\log n) \) a.s. for asymptotically efficient adaptive controllers.

2. Some preliminary lemmas. An important tool for the analysis of recursive identification and adaptive control algorithms is the following result from martingale theory.

**Lemma 1.** Let \( \{\xi_n\} \) be a martingale difference sequence with respect to an increasing sequence of \( \sigma \)-fields \( \{\mathcal{F}_n\} \) such that (1.3) holds. Let \( z_n \) be an \( \mathcal{F}_{n-1} \)-measurable random variable for every \( n \).

(i) \( \sum_{i=1}^{n} z_i w_i \) converges a.s. on \( \{\sum_{i=1}^{\infty} z_i^2 < \infty\} \), and for every \( \eta > 1/2 \),

\[
\left( \sum_{i=1}^{n} z_i w_i \right) / \left( \sum_{i=1}^{n} z_i^2 \right)^{\eta} \to 0 \quad \text{a.s. on } \{\sum_{i=1}^{\infty} z_i^2 = \infty\}.
\]

Consequently,

\[
(2.1) \quad \sum_{i=1}^{n} z_i w_i = o \left( \sum_{i=1}^{n} z_i^2 \right) + O(1) \quad \text{a.s.}
\]

(ii) \( \sum_{i=1}^{n} |z_i| w_i^2 = O(\sum_{i=1}^{n} |z_i|) \) a.s. on \( \{\sup_n |z_n| < \infty\} \). Moreover,

\[
(2.2) \quad \sum_{i=1}^{n} |z_i| w_i^2 = \sum_{i=1}^{n} |z_i| E(w_i^2 \mid \mathcal{F}_{i-1}) + o \left( \sum_{i=1}^{n} |z_i| \right) \quad \text{on } \{\sup_n |z_n| < \infty, \sum_{i=1}^{\infty} |z_n| = \infty\}.
\]

(iii) Let \( T_1 < T_2 < \cdots \) be a sequence of stopping times (with respect to \( \{\mathcal{F}_n\} \)) such that \( T_{j+1} - T_j \geq d (\geq 1) \). Let \( G_j \) be the \( \sigma \)-field generated by \( \mathcal{F}_{T_j+1} \) and let \( w_j = |\sum_{i=1}^{d} \alpha_i \varepsilon_{T_j+i} - E(\sum_{i=1}^{d} \alpha_i \varepsilon_{T_j+i} \mid \mathcal{F}_{T_j}) \), where \( \alpha_1, \cdots, \alpha_d \) are constants. Then \( \{w_j, G_j, j \geq 1\} \) is a martingale difference sequence with \( \sup_j E(|w_j|^{\alpha} \mid G_{j-1}) < \infty \) a.s.

For the proof of parts (i) and (ii) of the Lemma 1, see [19, p.157], while part (iii) follows from that \( w_j \) is \( G_j \)-measurable and that \( E(w_j \mid \mathcal{F}_{T_j}) = 0 \) and

\[
E(|w_j|^{\alpha} \mid \mathcal{F}_{T_j}) \leq 2^{\alpha} \sum_{k=1}^{\infty} I_{\{T_j=k\}} E(|\sum_{i=1}^{d} \alpha_i \varepsilon_{k+i}|^{\alpha} \mid \mathcal{F}_k).
\]

While Lemma 1 is probabilistic in nature, Lemmas 2–5 below are algebraic. In particular, the algebraic identity (2.3) in Lemma 2 will be applied to analyze various recursive identification
algorithms, in both the explicit and the implicit models. For a \( \nu \times \nu \) matrix \( A \), define \( \| A \| = \sup_{\| x \| = 1} \| Ax \| = \lambda_{\max}^{1/2}(A' A) \).

**Lemma 2.** Suppose that for \( n \leq t < m \), \( \phi_t = (y_t, \cdots, y_{t-\nu+1}, u_t, \cdots, u_{t-h-k+1}, \tilde{y}_{t+d}, \cdots, \tilde{y}_{t+d-h})' \) and \( C(q^{-1})(y_{t+d} - \eta_{t+d}) = G(q^{-1})y_t + q^{-\delta} \Gamma(q^{-1}) u_t \), where \( \delta \geq 0, d \geq 1, \nu \geq 1, \kappa \geq 1 \), and \( C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_h q^{-h} \), \( G(q^{-1}) = g_1 + \cdots + g_\nu q^{-(\nu-1)} \) and \( \Gamma(q^{-1}) = \gamma_1 + \cdots + \gamma_\kappa q^{-(\kappa-1)} \) are polynomials in the backward shift operator \( q^{-1} \). Suppose that there exist \( (\nu + \kappa + h) \times 1 \) vectors \( \theta_t \) such that \( \tilde{y}_{t+d} = \theta_t \phi_t \) for \( n \leq t < m \). Then

\[
(2.3) \quad C(q^{-1})(y_s - \tilde{y}_s - \eta_s) = -(\theta_{s-d} - \theta_t)' \phi_{s-d} \quad \text{for} \quad m + d > s \geq n + d,
\]

where \( \theta = (g_1, \cdots, g_\nu, \gamma_1, \cdots, \gamma_\kappa, -c_1, \cdots, -c_h)' \).

**Proof.** For \( n \leq t < m \),

\[
C(q^{-1})(y_{t+d} - \tilde{y}_{t+d} - \eta_{t+d}) = C(q^{-1})(y_{t+d} - \eta_{t+d}) - (C(q^{-1}) - 1) \tilde{y}_{t+d} - \tilde{y}_{t+d} \\
= (G(q^{-1})y_t + \Gamma(q^{-1}) u_t - \delta) - (c_1 y_{t+1} + \cdots + c_h y_{t+d-h} - \theta_t' \phi_t) \\
= \theta_t' \phi_t - \theta_t' \phi_t. \quad \Box
\]

**Lemma 3.** Let \( \{D_n, n \geq 0\} \) be a sequence of \( L \times L \) real matrices such that \( \sum_{n=0}^{\infty} \| D_n \| < \infty \) and let \( D(z) = \sum_{n=0}^{\infty} D_n z^n \). Suppose that \( D(e^{it}) + D'(e^{-it}) \) is nonnegative definite for all \( t \in [-\pi, \pi] \).

(i) Let \( \{g_n, n \geq 0\} \) be a sequence of \( L \times 1 \) real vectors and let \( f_n = \sum_{k=0}^{n} D_k g_{n-k} \). Then for any \( N \geq 0 \), \( \sum_{n=0}^{N} f_n' g_n \geq 0 \).

(ii) Suppose that \( D_i = 0 \) for all \( i > h \). Let \( M \geq h \) and suppose that \( f_n = \sum_{j=0}^{h} D_j g_{n-j} \) for \( M \leq n \leq N \). Let \( \{r_n, M \leq n \leq N\} \) be a nondecreasing sequence of positive numbers. Then

\[
\sum_{n=M}^{N} f_n' g_n / r_n \geq \sum_{j=1}^{h-j} \sum_{t=0}^{h-1} g_{M-1-t} D_{j+t} g_{M-1+j} / r_{M-1+j}.
\]

**Proof.** To prove (i), note that

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} \left( \sum_{n=0}^{N} g_n e^{-int} \right)' \left( \frac{D(e^{it}) + D'(e^{-it})}{2} \right) \left( \sum_{n=0}^{N} g_n e^{int} \right) dt \\
= \frac{1}{2\pi} \sum_{k=0}^{\infty} \sum_{n=0}^{N} \sum_{m=0}^{N} g_n' D_k g_m \int_{-\pi}^{\pi} e^{i(m-n)t} e^{ikt} dt \\
= \sum_{k=0}^{N} \sum_{m-n \leq k} g_n' D_k g_m = \sum_{n=0}^{N} g_n' \sum_{k=0}^{n} D_k g_{n-k} = \sum_{n=0}^{N} g_n' f_n.
\]
Since \( D(ε^{it}) + D'(ε^{-it}) \) is nonnegative definite, (i) follows.

To prove (ii), let \( g^n_\ast = g_n \) if \( n \geq M \) and let \( g^n_\ast = 0 \) otherwise. Let \( f^n_\ast = \sum_{j=0}^{h} D_j g^n_{n-j} \).

Then \( f^n_\ast = f_n \) for \( n \geq M + h \) and \( f_n - f^n_\ast = \sum_{m=n-h}^{M-1} D_{n-m} g_m \) for \( M \leq n < M + h \). Therefore

\[
\sum_{n=M}^{N} \frac{g^n_\ast f_n}{r_n} = \sum_{n=M}^{N} \frac{g_n^\ast f_n}{r_n} + \sum_{j=1}^{h} \sum_{k=j}^{h} \frac{D_k g_{M-1+j-k}}{r_{M-1+j}}.
\]

Let \( S_n = \sum_{m=0}^{n} g_m^\ast f_m \). By (i), \( S_n \geq 0 \) for all \( n \leq N \). Since \( g_n^\ast = 0 \) for \( n < M \), \( S_{M-1} = 0 \). Summation by parts gives

\[
\sum_{n=M}^{N} \frac{g_n^\ast f_n}{r_n} = \frac{1}{N-1} S_N + \sum_{n=M}^{N-1} \left( \frac{1}{r_n} - \frac{1}{r_{n+1}} \right) S_n \geq 0,
\]

noting that \( \frac{1}{r_n} \geq \frac{1}{r_{n+1}} > 0 \). Hence the desired conclusion follows. ■

**Lemma 4** Let \( x_1, x_2, \cdots \) be \( \nu \times 1 \) vectors and let \( A_n = A_{n-1} + x_n x_n' + \rho_n I \), where \( \rho_n \) are nonnegative scalars, \( I \) is the \( \nu \times \nu \) identity matrix and \( A_0 \) is a symmetric, positive definite \( \nu \times \nu \) matrix.

(i) If \( \lim_{n \to \infty} \lambda_{\max}(A_n) < \infty \), then \( \sum_{i=1}^{\infty} x_i' A_i^{-1} x_i < \infty \). If \( \lim_{n \to \infty} \lambda_{\max}(A_n) = \infty \), then

\[
(2.4) \quad \sum_{i=1}^{n} x_i' A_i^{-1} x_i \leq (1 + o(1)) \log \det A_n.
\]

(ii) Suppose that \( \sup_n \rho_n < \infty \) and that \( \lambda_{\min}(A_n) \to \infty \) and \( x_n' A_{n-1}^{-1} x_n \to 0 \) as \( n \to \infty \). Then for every fixed \( r = 0, \pm 1, \pm 2, \cdots \),

\[
(2.5) \quad x_n' A_{n+r}^{-1} x_n \sim x_n' A_{n}^{-1} x_n \quad \text{as} \quad n \to \infty.
\]

**Proof.** By Lemma 2(i) of [19],

\[
(2.6) \quad x_i' A_i^{-1} x_i = (|A_i| - |A_{i-1} + \rho_i I|)/|A_i|.
\]

Since \( |A_{i-1} + \rho_i I| \geq |A_{i-1}| \), it follows from (2.6) that \( x_i' A_i^{-1} x_i \leq (|A_i| - |A_{i-1}|)/|A_i| \) and therefore (i) follows by the same argument as that used to prove Lemma 2(ii) of [19].

To prove (ii), we first show that

\[
(2.7) \quad x_n' A_{n-1}^{-1} x_n \sim x_n' A_{n}^{-1} x_n \quad \text{as} \quad n \to \infty.
\]

By the matrix inversion lemma (cf. [3, p.824]),

\[
A_n^{-1} = (A_{n-1} + \rho_n I)^{-1} - \frac{(A_{n-1} + \rho_n I)^{-1} x_n x_n' (A_{n-1} + \rho_n I)^{-1}}{1 + x_n' (A_{n-1} + \rho_n I)^{-1} x_n},
\]

where
and therefore

\begin{equation}
(2.8) \quad x_n' A_n^{-1} x_n = x_n' (A_{n-1} + \rho_n I)^{-1} x_n / \{1 + x_n' (A_{n-1} + \rho_n I)^{-1} x_n\}.
\end{equation}

Since \((A_{n-1} + \rho_n I)^{-1} = A_{n-1}^{-1/2} (I + \rho_n A_{n-1}^{-1})^{-1} A_{n-1}^{-1/2} = A_{n-1}^{-1} + A_{n-1}^{-1/2} B_n A_{n-1}^{-1/2}\), where \(B_n = \sum_{i=1}^{\infty} (-\rho_n A_{n-1}^{-1})^i\), and since \(\rho_n \lambda_{\max}(A_{n-1}^{-1}) \to 0\) (implying that \(\lambda_{\max}(B_n) \to 0\)), it then follows that

\begin{equation}
(2.9) \quad x_n' (A_{n-1} + \rho_n I)^{-1} x_n = x_n' A_{n-1}^{-1} x_n + o(||A_{n-1}^{-1/2} x_n||^2) = (1 + o(1)) x_n' A_{n-1}^{-1} x_n.
\end{equation}

From (2.8) and (2.9), (2.7) follows.

We next show by induction that for \(r = 1, 2, \ldots\),

\begin{equation}
(2.10) \quad x_n' A_{n-r}^{-1} x_n \sim x_n' A_{n-1}^{-1} x_n \quad \text{as} \quad n \to \infty.
\end{equation}

Note that (2.10) reduces to (2.7) when \(r = 1\). Suppose that (2.10) holds for \(1 \leq r \leq s - 1\). Since \(A_{n-1} = A_{n-s} + \sum_{j=1}^{s-1} (x_{n-j} A_n^{-1} x_{n-j} + \rho_{n-j} I)\),

\begin{equation}
(2.11) \quad A_{n-1}^{-1} = A_{n-s}^{-1} \{I + \sum_{j=1}^{s-1} (A_{n-s}^{-1/2} x_{n-s} x_{n-s-j} A_{n-s}^{-1/2} + \rho_{n-j} A_{n-s}^{-1})\}^{-1} A_{n-s}^{-1/2}
\end{equation}

\[= A_{n-s}^{-1} + A_{n-s}^{-1/2} C_n A_{n-s}^{-1/2},\]

where

\begin{equation}
(2.12) \quad C_n = \sum_{i=1}^{\infty} \left\{ - \sum_{j=1}^{s-1} (A_{n-s}^{-1/2} x_{n-s} x_{n-s+j} A_{n-s}^{-1/2} + \rho_{n-s+j} A_{n-s}^{-1}) \right\}^i.
\end{equation}

Noting that \(\|C\| = \lambda_{\max}(C) \leq \text{tr}(C)\) if \(C\) is symmetric and nonnegative definite, we have

\begin{equation}
(2.13) \quad \sum_{j=1}^{s-1} ||A_{n-s}^{-1/2} x_{n-s+j} x_{n-s+j} A_{n-s}^{-1/2}|| \leq \sum_{j=1}^{s-1} x_{n-s+j} A_{n-s}^{-1} x_{n-s+j} \to 0,
\end{equation}

since (2.10) holds for \(1 \leq r \leq s - 1\). From (2.13) and the fact that \(\rho_{n-s+j} \lambda_{\max}(A_{n-s}^{-1}) = \rho_{n-s+j} / \lambda_{\min}(A_{n-s}) \to 0\), it follows that \(\|C_n\| \to 0\), and therefore by (2.11),

\begin{equation}
(2.14) \quad x_n' A_{n-1}^{-1} x_n = x_n' A_{n-s}^{-1} x_n + o(x_n' A_{n-s}^{-1} x_n).
\end{equation}

In view of (2.14) and (2.7), (2.10) holds for \(r = s\).

For \(r \geq 1\), since \(A_{n+r} = A_n + \sum_{j=1}^{r} (x_{n+j} x_{n+j} + \rho_{n+j} I)\), we can make use of (2.10) and the same argument as in (2.11)-(2.14) to show that \(x_n' A_{n+r}^{-1} x_n = x_n' A_{n-1}^{-1} x_n + o(x_n' A_{n-1}^{-1} x_n)\). 

11
LEMMA 5. Suppose that the polynomial $C(z) = 1 + c_1z + \cdots + c_hz^h$ is stable. For $j = 1, \ldots, h$, let $\{c_{n,j}\}$ be a sequence of numbers such that $\lim_{n \to \infty} c_{n,j} = c_j$, and let $C_n(q^{-1}) = 1 + c_{n,1}q^{-1} + \cdots + c_{n,h}q^{-h}$.

(i) Suppose that $\xi_n$ and $\phi_n$ are $n \times 1$ vectors such that $C_n(q^{-1})\xi_n = \phi_n$. Then there exist $K > 0$ and $0 < \rho < 1$ such that for all $t > m$

$$
\|\xi_t\| \leq K\left\{\sum_{i=0}^{t-m-1} \rho^i\|\phi_{t-i}\| + \rho^{t-m}\sum_{r=0}^{h-1}\|\xi_{m-r}\|\right\}.
$$

Consequently, there exists $K' > 0$ such that for all $n > m$,

$$
\sum_{t=m+1}^{n} \|\xi_t\|^2 \leq K'\left\{\max_{0 \leq r \leq h-1}\|\xi_{m-r}\|^2 + \sum_{t=m+1}^{n}\|\phi_t\|^2\right\}.
$$

(ii) Let $G(q^{-1}) = g_1 + \cdots + g_pq^{-p+1}$, $\Gamma(q^{-1}) = \gamma_1 + \cdots + \gamma_kq^{-k+1}$, $G_n(q^{-1}) = g_{n,1} + \cdots + g_{n,p}q^{-p+1}$, $\Gamma_n(q^{-1}) = \gamma_{n,1} + \cdots + \gamma_{n,k}q^{-k+1}$, where $\lim_{n \to \infty} g_{n,j} = g_j$ and $\lim_{n \to \infty} \gamma_{n,j} = \gamma_j$ for every $j$. Suppose that

$$
C(q^{-1})\tilde{y}_{n+d} = G(q^{-1})y_n + \Gamma(q^{-1})u_n, \quad C_n(q^{-1})\tilde{y}_{n+d} = G_n(q^{-1})y_n + \Gamma_n(q^{-1})u_n.
$$

Then $\sum_{i=1}^{n} (\tilde{y}_{i+d} - \tilde{y}_{i+d})^2 = o\left(\sum_{i=1}^{n} (y_i^2 + u_i^2)\right)$.

Proof. For (i), see [19, pp. 161-162] and [9, p.904]. To prove (ii), note that by (2.17),

$$
C_n(q^{-1})(\tilde{y}_{n+d} - \tilde{y}_{n+d}) = C_n(q^{-1})\tilde{y}_{n+d} - C(q^{-1})\tilde{y}_{n+d} - (C_n(q^{-1}) - C(q^{-1}))\tilde{y}_{n+d}
$$

$$
= (G_n(q^{-1}) - G(q^{-1}))y_n + (\Gamma_n(q^{-1}) - \Gamma(q^{-1}))u_n - (C_n(q^{-1}) - C(q^{-1}))\tilde{y}_{n+d}.
$$

Since $G_n - G \to 0$, $\Gamma_n - \Gamma \to 0$ and $C_n - C \to 0$, it follows from (2.18) and part (i) of the lemma that

$$
\sum_{i=1}^{n} (\tilde{y}_{i+d} - \tilde{y}_{i+d})^2 = o\left(\sum_{i=1}^{n} (y_i^2 + u_i^2 + \tilde{y}_{i+d}^2)\right).
$$

Again by (2.17) and (i),

$$
\sum_{i=1}^{n} \tilde{y}_{i+d}^2 = O\left(\sum_{i=1}^{n} (y_i^2 + u_i^2)\right).
$$

From (2.19) and (2.20), the desired conclusion follows.

3. The stochastic gradient algorithm and some extensions. To begin with, consider the stochastic gradient algorithm (1.30) that estimates the parameter vector (1.11) of
the explicit system (1.1). As an application of Lemma 1 to the extended stochastic Lyapunov function \( Q_n = \|\Theta_n - \Theta\|^2 \), we prove the following.

**COROLLARY 1.** Suppose that \( \min_{|z| = 1} \text{Re}\{C(z) - a/2\} > 0 \) and that the random disturbances \( \epsilon_n \) in the linear stochastic system (1.1) satisfy (1.3).

(i) For the stochastic gradient algorithm (1.30),

\[
\limsup_{n \to \infty} \|\Theta_n\| < \infty \quad \text{a.s.},
\]

\[
\sum_{n=1}^{\infty} (\epsilon_n - \epsilon_n)^2 / r_n < \infty \quad \text{a.s.},
\]

\[
\sum_{n=1}^{\infty} \|\Theta_n - \Theta_{n-1}\|^2 < \infty \quad \text{a.s.},
\]

where \( \epsilon_n \) and \( r_n \) are defined in (1.30b) and (1.30c).

(ii) Let \( \{D_n\} \) be a sequence of closed convex regions in \( R^{p+k+h} \) such that \( \Theta \in D_n \) for all large \( n \). Let \( \Pi_{D}(x) \) denote the Euclidean projection of \( x \) into \( D \), i.e., \( \|x - \Pi_{D}(x)\| = \min\{\|x - y\| : y \in D\} \). Suppose that we modify (1.30a) defining the stochastic gradient algorithm as

\[
\Theta_n = \Pi_{D_n}(\Theta_{n-1} + a r_{n-1}^{-1} \epsilon_n \Phi_{n-1}).
\]

Then (3.1) and (3.2) still hold. If furthermore \( D_n \supset D_{n-1} \) for all large \( n \), then (3.3) still holds.

**Proof.**

(i) Let \( Q_i = \|\Theta_i - \Theta\|^2 \). From (1.30a) it follows that

\[
Q_i = Q_{i-1} + 2a r_{i-1}^{-1} \epsilon_i \Phi_{i-1}(\Theta_{i-1} - \Theta) + a^2 r_{i-1}^{-2} \epsilon_i^2 \|\Phi_{i-1}\|^2.
\]

Replacing \( \epsilon_i \) in (3.5) by \( (\epsilon_i - \epsilon_i) + \epsilon_i \) and summing (3.5) over \( i \) from \( h+1 \) to \( n \), we obtain that for \( n > h \),

\[
Q_n = Q_{h} + 2a \sum_{i=h+1}^{n} \{r_{i-1}^{-1}(\epsilon_i - \epsilon_i)(\Theta_{i-1} - \Theta) + (a/2) r_{i-1}^{-2}(\epsilon_i - \epsilon_i)^2 \|\Phi_{i-1}\|^2\}
\]

\[
+ a^2 \sum_{i=h+1}^{n} r_{i-1}^{-2} \|\Phi_{i-1}\|^2 \epsilon_i^2 + 2a \sum_{i=h+1}^{n} \epsilon_i \{r_{i-1}^{-1} \Phi_{i-1}(\Theta_{i-1} - \Theta) + a r_{i-1}^{-2}(\epsilon_i - \epsilon_i) \|\Phi_{i-1}\|^2\}.
\]

Since \( \epsilon_i - \epsilon_i = (y_i - \epsilon_i) - (\Theta_{i-1} - \Theta) \Phi_{i-1} \) is \( F_{i-1} \)-measurable, it follows from (2.1) that

\[
\sum_{i=h+1}^{n} \epsilon_i \{r_{i-1}^{-1} \Phi_{i-1}(\Theta_{i-1} - \Theta) + a r_{i-1}^{-2}(\epsilon_i - \epsilon_i) \|\Phi_{i-1}\|^2\}
\]

\[
= o(\sum_{i=h+1}^{n} r_{i-1}^{-2} (\Theta_{i-1} - \Theta)^2) + o(\sum_{i=h+1}^{n} r_{i-1}^{-2}(\epsilon_i - \epsilon_i)^2) + O(1) \quad \text{a.s.},
\]

13
noting that $\|\Phi_t\|^2/r_t \leq 1$. Moreover, since $\sum_{i=1}^{\infty} r_{i-1}^{-2} \|\Phi_t\|^2 < \infty$, it follows from Lemma 1(ii) that

$$
(3.8) \quad \sum_{i=h+1}^{n} r_{i-1}^{-2} \|\Phi_{i-1}\|^2 e_i^2 = O\left( \sum_{i=h+1}^{n} r_{i-1}^{-2} \|\Phi_{i-1}\|^2 \right) = O(1) \quad \text{a.s.}
$$

By continuity and compactness, we can choose $\rho > 0$ such that $\min_{|z|=1} \text{Re}\{C(z) - (a + \rho)/2\} > 0$. Moreover, by Lemma 2 (with $d = 1$ and $\eta_t = \epsilon_t$),

$$
(3.9) \quad \Phi_{i-1}'(\Theta_{i-1} - \Theta) + (a + \rho)(e_i - \epsilon_i)/2 = -\left\{ C(q^{-1}) - \frac{1}{2}(a + \rho) \right\} (e_i - \epsilon_i).
$$

Since $\|\Phi_t\|^2/r_t \leq 1$, it then follows from Lemma 3(ii) that

$$
(3.10) \quad \sum_{i=h+1}^{n} \left\{ r_{i-1}^{-1} (e_i - \epsilon_i) \Phi_{i-1}'(\Theta_{i-1} - \Theta) + (a/2) r_{i-1}^{-2} (e_i - \epsilon_i)^2 \|\Phi_{i-1}\|^2 \right\}
$$

$$
\leq \sum_{i=h+1}^{n} r_{i-1}^{-1} (e_i - \epsilon_i) \left\{ \Phi_{i-1}'(\Theta_{i-1} - \Theta) + (a + \rho)(e_i - \epsilon_i)/2 \right\} - (\rho/2) \sum_{i=h+1}^{n} r_{i-1}^{-1} (e_i - \epsilon_i)^2
$$

$$
\leq -\left( \frac{\rho}{2} \right) \sum_{i=h+1}^{n} r_{i-1}^{-1} (e_i - \epsilon_i)^2 + O(1).
$$

Since $\Phi_{i-1}'(\Theta_{i-1} - \Theta) = -C(q^{-1})(e_i - \epsilon_i)$,

$$
(3.11) \quad \sum_{i=h+1}^{n} r_{i-1}^{-2} [\Phi_{i-1}'(\Theta_{i-1} - \Theta)]^2 = O\left( \sum_{i=1}^{n} r_{i-1}^{-2} (e_i - \epsilon_i)^2 \right).
$$

From (3.6)-(3.11), it follows that $Q_n \leq \{-a \rho + o(1)\} \sum_{i=1}^{n} r_{i-1}^{-1} (e_i - \epsilon_i)^2 + O(1) \text{a.s.}$, giving the desired conclusion (3.1) and (3.2). To prove (3.3), note that by (1.30a),

$$
\|\Theta_i - \Theta_{i-1}\|^2 = a^2 r_{i-1}^{-2} \|\Phi_{i-1}\|^2 \{e_i^2 + 2a(e_i - \epsilon_i) + (e_i - \epsilon_i)^2\}.
$$

Therefore by (3.7) and (3.8),

$$
\sum_{i=1}^{n} \|\Theta_i - \Theta_{i-1}\|^2 \leq (a^2 + o(1)) \sum_{i=1}^{n} r_{i-1}^{-1} (e_i - \epsilon_i)^2 + O(1) \quad \text{a.s.,}
$$

noting that $\|\Phi_t\|^2/r_t \leq 1$. Hence (3.3) follows from (3.2).

(ii) Suppose that $\Theta \in D_i$ for all $i \geq m(> h)$. For $i \geq m$, since $D_i$ is convex and $\Theta \in D_i$, it follows from (3.4) that

$$
\|\Theta_i - \Theta_{i-1}\|^2 \leq \|\Theta_{i-1} + ar_{i-1}^{-1} e_i \Phi_{i-1} - \Theta\|^2
$$

$$
= \|\Theta_{i-1} - \Theta\|^2 + 2ar_{i-1}^{-1} e_i \Phi_{i-1} (\Theta_{i-1} - \Theta) + a^2 r_{i-1}^{-2} e_i^2 \|\Phi_{i-1}\|^2.
$$
and therefore the same argument as before proves that (3.1) and (3.2) still hold. Suppose that \( D_{i-1} \subset D_i \) for all \( i \geq m \). Then for \( i \geq m \), since \( \Theta_{i-1} \in D_{i-1} \subset D_i \) and \( D_i \) is convex, it follows from (3.4) that

\[
\|\Theta_i - \Theta_{i-1}\|^2 \leq \| (\Theta_{i-1} + a r_{i-1}^{-1} e_i \Phi_{i-1}) - \Theta_i \|^2 = a^2 r_{i-1}^{-2} \| \Phi_{i-1} \|^2 e_i^2,
\]

and the same argument as before shows that (3.3) still holds. \( \blacksquare \)

The preceding proof uses the martingale limit theorems in Lemma 1 to analyze directly the recursive inequalities for \( Q_n = \| \Theta_n - \Theta \|^2 \). In contrast, the method of stochastic Lyapunov functions used by Goodwin et al. [11], [14], [15] to analyze the stochastic gradient algorithm, under the additional assumption that \( C(z) \) is stable, relies on the martingale convergence theorem applied to the transformation \( Z_n = Q_n + S_n/r_{n-1} \) with \( S_n \geq 0 \). As shown in the proof of Theorem 1(ii), these recursive inequalities also apply to the modified stochastic gradient algorithm (3.4) that constrains by projection the estimator to lie inside some convex region. Such constrained algorithms have been studied by Ljung [5], [6] and Kushner and Clark [20] by the ODE method. Making use of (3.1)-(3.3), Fuchs [16], [17] established the global convergence property (1.22) for the adaptive d-step ahead predictors defined by the explicit approach (1.19) in which the estimated polynomials are given by the stochastic gradient algorithm \( \Theta_n \), under the assumptions (1.21), (1.26) and (1.27).

The ARMAX model (1.1) can be written as a linear regression model (1.12) and (1.30) represents the stochastic gradient algorithm to estimate the parameter vector \( \Theta \) of this regression model. For d-step ahead prediction, we can express (1.1) as an alternative regression model (1.9), in which \( E(\eta_{n+d}|F_n) = 0 \) while \( \psi_n \) is \( F_n \)-measurable. The stochastic gradient algorithm estimating the parameter \( \theta \) in the regression model \( y_n = \theta^t \psi_{n-d} + \eta_n \) (i.e., (1.9)) takes the form

\[
\begin{align*}
\theta_n &= \theta_{n-1} + (a/r_{n-d}) \phi_{n-d}(y_n - \hat{y}_n), \\
\phi_n &= (y_n, \ldots, y_{n-p(d)+1}, u_{n-\Delta+d}, \ldots, u_{n-\Delta-k+2}, \hat{y}_{n+d-1}, \ldots, \hat{y}_{n+d-k}), \\
r_n &= r_{n-1} + \| \phi_n \|^2, \\
\hat{y}_{n+d} &= \theta_n^t \phi_n,
\end{align*}
\]

where \( p(d) = p \vee (h - d + 1) \). The extended stochastic Lyapunov function argument used in Corollary 1 can be modified to prove the global convergence of the adaptive predictors (3.12d) based on the stochastic gradient algorithm that does not involve interlacing in the implicit approach. This is the content of

**THEOREM 1.** Suppose that \( \min_{|z| = 1} \text{Re\{C(z) - (d - 1/2)a\}} > 0 \) and that the random disturbances \( \eta_n \) in the linear stochastic system (1.1) satisfy (1.3). Consider the stochastic gradient algorithm \( \theta_n \) defined by (3.12) for the implicit model (1.9) in which the \( \hat{y}_i \) and \( \eta_i \) are
defined by (1.6) and (1.7). Then

\begin{equation}
\limsup_{n \to \infty} \|\theta_n\| < \infty \text{ a.s., } \sum_{d+1}^{\infty} \|\theta_n - \theta_{n-1}\|^2 < \infty \text{ a.s.,}
\end{equation}

(3.13)

\begin{equation}
\sum_{n=d+1}^{\infty} \left( \tilde{y}_n - \tilde{y}_n \right)^2/ r_{n-d} < \infty \text{ a.s.}
\end{equation}

(3.14)

Consequently

\begin{equation}
n^{-1} \sum_{t=d+1}^{n} (\tilde{y}_t - \tilde{y}_t)^2 \to 0 \text{ a.s. on } \{ \limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} (y_i^2 + u_i^2) < \infty \}.
\end{equation}

(3.15)

If \(B(z) \neq 0\) and is stable, then on \(\{ \limsup_{n \to \infty} n^{-1} \sum_{t=1}^{n} y_t^2 < \infty \} \cup \{ \limsup_{n \to \infty} n^{-1} \sum_{t=1}^{n} \tilde{y}_t^2 < \infty \},\)

\begin{equation}
n^{-1} \sum_{t=d+1}^{n} (\tilde{y}_t - \tilde{y}_t)^2 \to 0 \text{ and } \sum_{t=1}^{n} (y_t^2 + u_t^2 + \tilde{y}_t^2) = O(n) \text{ a.s.}
\end{equation}

(3.16)

If \(A(z)\) is stable, then (3.16) holds on \(\{ \limsup_{n \to \infty} n^{-1} \sum_{t=1}^{n} u_t^2 < \infty \}).\)

Proof. Let \(Q_n = \|\theta_n - \theta\|^2, \epsilon_n = y_n - \tilde{y}_n\). As in (3.6), it follows from (3.12a) that for \(n > m,\)

\begin{equation}
Q_n = Q_m + 2a \sum_{m+1}^{n} \left\{ r_{i-d}^{-1}(e_i - \eta_i) \phi_{i-d}(\theta_{i-1} - \theta) + (a/2)r_{i-d}^{-2}(e_i - \eta_i)^2 \|\phi_{i-d}\|^2 \right\}
\end{equation}

\begin{equation}
+ a^2 \sum_{m+1}^{n} r_{i-d}^{-2} \|\phi_{i-d}\|^2 \eta_i^2 + 2a \sum_{m+1}^{n} \eta_i \{ r_{i-d}^{-1} \phi_{i-d}(\theta_{i-1} - \theta) + ar_{i-d}^{-2}(e_i - \eta_i) \|\phi_{i-d}\|^2 \}.
\end{equation}

(3.17)

Note that \(r_{i-d}, \phi_{i-d}, \theta_{i-d}\) and \(e_i - \eta_i = \theta' \psi_{i-d} - \theta'_{i-d} \phi_{i-d}\) are \(\mathcal{F}_{i-d}\)-measurable. Since \(\eta_i = \epsilon_{i+1} + \cdots + \epsilon_{i-d+1}\), it follows from Lemma 1 that as \(n \to \infty,\)

\begin{equation}
\sum_{m+1}^{n} \eta_i \{ r_{i-d}^{-1} \phi_{i-d}(\theta_{i-d} - \theta) + ar_{i-d}^{-2}(e_i - \eta_i) \|\phi_{i-d}\|^2 \}
\end{equation}

\begin{equation}
= o \left( \sum_{m+1}^{n} r_{i-d}^{-2} \|\phi_{i-d}(\theta_{i-d} - \theta)\|^2 \right) + o \left( \sum_{m+1}^{n} r_{i-d}^{-2}(e_i - \eta_i)^2 \right) + O(1) \text{ a.s.,}
\end{equation}

(3.18)

\begin{equation}
\sum_{m+1}^{n} r_{i-d}^{-2} \|\phi_{i-d}\|^2 \eta_i^2 = O \left( \sum_{m+1}^{n} r_{i-d}^{-2} \|\phi_{i-d}\|^2 \right) = O(1) \text{ a.s.,}
\end{equation}

(3.19)

16
analogous to (3.7) and (3.8). Choosing $\rho > 0$ such that $\text{Re}\{C(z) - (d - 1/2)a - \rho\} > 0$ for all $|z| = 1$, we obtain by Lemma 2 that as in (3.9) and (3.10),

$$
(3.20) \quad \sum_{m+1}^{n} \{r_{i-d}^{-1}(e_i - \eta_i)\phi'_{i-d}(\theta_{i-d} - \theta) + (a/2)r_{i-d}^{-2}(e_i - \eta_i)^2\|\phi_{i-d}\|^2\}
$$

$$
\leq - \sum_{m+1}^{n} r_{i-d}^{-1}(e_i - \eta_i) \{[C(q^{-1}) - a/2 - \rho](e_i - \eta_i)\} - \rho \sum_{m+1}^{n} r_{i-d}^{-1}(e_i - \eta_i)^2.
$$

Noting that (3.17) involves $\theta_{i-1} - \theta$ instead of $\theta_{i-d} - \theta$, we write in the case $d > 1$

$$
(3.21) \quad \theta_{i-1} - \theta = (\theta_{i-d} - \theta) + \sum_{s=1}^{d-1}(\theta_{i-s} - \theta_{i-s-1}).
$$

Fix $s = 1, \cdots, d - 1$. From (3.12a) and the inequality $|x'y| \leq (\|x\|^2 + \|y\|^2)/2$, it follows that

$$
(3.22) \quad \left| \sum_{m+1}^{n} r_{i-d}^{-1}(e_i - \eta_i)\phi'_{i-d}(\theta_{i-s} - \theta_{i-s-1}) \right|
$$

$$
= \left| a \sum_{m+1}^{n} [r_{i-d}^{-1}(e_i - \eta_i)\phi_{i-d}]'[r_{i-d}^{-1}e_i - \phi_{i-d}] \right|
$$

$$
\leq (a/2)\{ \sum_{m+1}^{n} r_{i-d}^{-2}(e_i - \eta_i)^2\|\phi_{i-d}\|^2 + \sum_{m+1}^{n} r_{i-d}^{-2}\|\phi_{i-d}\|^2 \}
$$

$$
= (a + o(1)) \sum_{m+1}^{n} r_{i-d}^{-2}(e_i - \eta_i)^2\|\phi_{i-d}\|^2 + O(1) \quad \text{a.s.},
$$

where the last relation above follows from an application of Lemma 1 to

$$
\sum_{m+1}^{n} r_{i-s-d}^{-2}e_i^2\|\phi_{i-s-d}\|^2 = \sum_{m+1}^{n} r_{i-s-d}^{-2}(e_i - \eta_i)^2\|\phi_{i-s-d}\|^2
$$

$$
+ \sum_{m+1}^{n} r_{i-s-d}\|\phi_{i-s-d}\|^2\eta_i^2 + 2 \sum_{m+1}^{n} r_{i-s-d}\|\phi_{i-s-d}\|^2(e_i - \eta_i)\eta_{i-s},
$$

noting that $e_i - \eta_i$ is $F_{i-s-d}$-measurable and that $\|\phi_t\|^2/r_t \leq 1$ and $\sum_{1}^{\infty} \|\phi_t\|^2/r_t^2 < \infty$. Moreover, analogous to (3.22), we have

$$
(3.23) \quad \left| \sum_{m+1}^{n} \eta_i r_{i-d}^{-1}\phi'_{i-d}(\theta_{i-s} - \theta_{i-s-1}) \right|
$$

17
\begin{align*}
&\leq a \left[ \sum_{m+1}^{n} \eta_{i} r_{i-d}^{-1} r_{i-s-d}^{-1} \phi_{i-d} \phi_{i-s-d} (e_{i-s} - \eta_{i-s}) \right] \\
&\quad + a \left[ \sum_{m+1}^{n} [\eta_{i} r_{i-d}^{-1} \phi_{i-d}] [\eta_{i-s} r_{i-s-d}^{-1} \phi_{i-s-d}] \right] \\
&= \eta( \sum_{m+1}^{n} r_{i-s-d}^{-1} (e_{i-s} - \eta_{i-s})^2 ) + O(1) + O( \sum_{m+1}^{n} r_{i-d}^{-2} \| \phi_{i-d} \|^2 \eta_i^2 ) \\
&= o( \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i})^2 ) + O(1) \quad \text{a.s., by Lemma 1.}
\end{align*}

Combining (3.21) with (3.20) and (3.22) gives

\begin{equation}
\sum_{m+1}^{n} \{ r_{i-d}^{-1} (e_{i} - \eta_{i}) \phi_{i-d} (\theta_{i-1} - \theta) + (a/2) r_{i-d}^{-2} (e_{i} - \eta_{i})^2 \| \phi_{i-d} \|^2 \}
\end{equation}

\begin{align*}
&\leq \left\{ (d - 1) a + o(1) \right\} \sum_{m+1}^{n} r_{i-d}^{-1} \| \phi_{i-d} \|^2 (e_{i} - \eta_{i})^2 + O(1) - \rho \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i})^2 \\
&\quad - \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i}) \{ C(q^{-1}) - a/2 - \rho \}(e_{i} - \eta_{i}) \\
&\leq -(\rho + o(1)) \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i})^2 + O(1) - \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i}) \{ C(q^{-1}) - (d - 1/2) a - \rho \}(e_{i} - \eta_{i}) \\
&\leq -(\rho + o(1)) \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i})^2 + O(1),
\end{align*}

where the last inequality follows from Lemma 3(ii). As in (3.11), it follows from Lemma 2 that

\begin{equation}
\sum_{m+1}^{n} r_{i-d}^{-2} [\phi_{i-d} (\theta_{i-d} - \theta)]^2 = O( \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i})^2 ).
\end{equation}

From (3.17)-(3.19) together with (3.21) and (3.23)-(3.25), it follows that $Q_n + (2 a \rho + o(1)) \sum_{m+1}^{n} r_{i-d}^{-1} (e_{i} - \eta_{i})^2 = O(1) \text{ a.s.},$ implying (3.13) and (3.14) as in the proof Corollary 1. Note in this connection that $e_{i} - \eta_{i} = y_i - \widehat{y}_i - \eta_{i} = \widehat{y}_i - \widehat{\eta}_i.$

By (3.14) and the Kronecker lemma,

\begin{equation}
\sum_{i=1}^{n} (y_i - \widehat{y}_i - \eta_{i})^2 = O(1) \quad \text{on } \{ \sup r_n < \infty \},
\end{equation}

\begin{align*}
&= o( \sum_{i=1}^{n-d} y_i^2 + \sum_{i=1}^{n-\Delta} u_i^2 + \sum_{i=1}^{n-1} \widehat{y}_i^2 ) \quad \text{on } \{ \sup r_n = \infty \}.
\end{align*}
From Lemma 1(ii), (3.26) and the inequality

\[(3.27) \quad \sum_{i=1}^{n} \tilde{y}_i^2 \leq 2 \sum_{i=1}^{n} (\tilde{y}_i + \eta_i - y_i)^2 + 4 \sum_{i=1}^{n} y_i^2 + 4 \sum_{i=1}^{n} \eta_i^2,\]

it follows that

\[(3.28) \quad \sum_{i=1}^{n} \tilde{y}_i^2 = O(n) \quad \text{a.s. on } \{\limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} (y_i^2 + u_i^2) < \infty\}.\]

From (3.26) and (3.28), (3.15) follows.

Suppose that \(B(z) \neq 0\) and is stable. By changing \(\Delta\) if necessary, we can assume that \(b_1 \neq 0\). Then by (1.1) and Lemma 5(i),

\[(3.29) \quad \sum_{i=1}^{n-\Delta} u_i^2 = O(\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \epsilon_i^2) = O(\sum_{i=1}^{n} y_i^2) + O(n) \quad \text{a.s.},\]

and therefore

\[(3.30) \quad \sum_{i=1}^{n} u_i^2 = O(n) \quad \text{a.s. on } \{\limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} y_i^2 < \infty\}.\]

From (3.26) and (3.29), it follows that

\[
\sum_{i=1}^{n} y_i^2 \leq 2 \sum_{i=1}^{n} (y_i - \tilde{y}_i - \eta_i)^2 + 4 \sum_{i=1}^{n} \tilde{y}_i^2 + 4 \sum_{i=1}^{n} \eta_i^2
\]

\[
= o\left(\sum_{i=1}^{n-\Delta} y_i^2\right) + O\left(\sum_{i=1}^{n} \tilde{y}_i^2\right) + O(n) \quad \text{a.s.,}
\]

and therefore

\[(3.31) \quad \sum_{i=1}^{n} y_i^2 = O(n) \quad \text{a.s. on } \{\limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} \tilde{y}_i^2 < \infty\}.\]

On the event \(\{\limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} y_i^2 < \infty\} \cup \{\limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} \tilde{y}_i^2 < \infty\}\), it follows from (3.28), (3.30) and (3.31) that \(\sum_{i=1}^{n} (y_i^2 + u_i^2 + \tilde{y}_i^2) = O(n)\) a.s., and therefore \(\sum_{i=1}^{n} (y_i - \tilde{y}_i - \eta_i)^2 = o(n)\) a.s. by (3.26).

Now assume that \(A(z)\) is stable. Then by (1.1) and Lemma 5(i),

\[
\sum_{i=1}^{n} y_i^2 = O\left(\sum_{i=1}^{n-\Delta} u_i^2\right) + O\left(\sum_{i=1}^{n} \epsilon_i^2\right) = O\left(\sum_{i=1}^{n-\Delta} u_i^2\right) + O(n) \quad \text{a.s.}
\]
Hence on the event \( \{ \limsup_{n \to \infty} n^{-1} \sum_{i=1}^{n} u_i^2 < \infty \} \), \( \sum_{i=1}^{n} y_i^2 = O(n) \) a.s., and therefore \( \sum_{i=1}^{n} \tilde{y}_i^2 = O(n) \) a.s. by (3.28) and \( \sum_{i=1}^{n} (y_i - \tilde{y}_i - \eta_i)^2 = o(n) \) a.s. by (3.26).

4. Extended least squares, modified least squares, and the associated adaptive predictors. To begin with, consider the model (1.1) with \( C(q^{-1}) = 1 \), i.e., the ARX model. In this case, the AML algorithm (1.13) reduces to the usual least squares estimator, for which (1.13b) becomes \( \Phi_n = (y_n, \ldots, y_{n-p+1}, u_{n-D+1}, \ldots, u_{n-D-k+2})' = \Psi_n \). As shown in [18] and [19], the least squares 1-step ahead predictors \( \tilde{y}_{n+1} = \Theta'_n \Phi_n \) satisfy

\[
\sum_{n=1}^{N} (\tilde{y}_{n+1} - \tilde{y}_{n+1})^2 I_{\{x'_n P_n \Phi_n \leq \delta \}} = O(\log \det P_N^{-1}) \quad \text{for every } 0 < \delta < 1,
\]

where \( \tilde{y}_{n+1} = \Theta'_n \Psi_n \) is the optimal 1-step ahead predictor of \( y_{n+1} \) assuming knowledge of the parameter vector \( \Theta = (-a_1, \ldots, -a_p, b_1, \ldots, b_k)' \) and \( P_n^{-1} = P_{n-1}^{-1} + \Phi_n \Phi_n' \) as in (1.13a). When the sample mean square boundedness assumption (1.21) holds for the input-output data, (4.1) implies that

\[
\sum_{n=1}^{N} (\tilde{y}_{n+1} - \tilde{y}_{n+1})^2 = O(\log N) \quad \text{a.s. on } \limsup_{n \to \infty} \Phi_n' P_n \Phi_n < 1.
\]

For general \( C(z) \) satisfying the positive real condition (1.15), Lai and Wei [9] showed that (4.1) still holds for the adaptive 1-step ahead predictors \( \tilde{y}_{n+1} = \Theta'_n \Phi_n \) based on the AML algorithm \( \Theta_n \) and the pseudoregression vector \( \Phi_n \) defined in (1.13). It is also shown in [9] that

\[
\sum_{n=1}^{N} \| \Phi_n - \Psi_n \|^2 = O(\log(2 + \sum_{n=1}^{N} \| \Phi_n \|^2)) \quad \text{a.s.}
\]

If (1.21) holds, then since \( \sum_{i=1}^{n} e_i^2 = O(N) \) a.s. by Lemma 1(ii), \( \sum_{n=1}^{N} \| \Psi_n \|^2 = O(N) \) a.s. and therefore \( \sum_{n=1}^{N} \| \Phi_n \|^2 = O(N) \) a.s. by (4.3). Hence, under the assumptions (1.15) and (1.21), (4.2) still holds for the adaptive predictor \( \tilde{y}_{n+1} \) based on the AML algorithm.

In [9], (4.1) was obtained from an analysis of the recursive inequalities for the extended stochastic Lyapunov function (1.16) that was also used to study the consistency of \( \Theta_n \). Earlier Solo [8] used these recursive inequalities to transform (1.16) into a nonnegative almost supermartingale and thereby applies the martingale convergence theorem to establish the strong consistency of \( \Theta_n \) under the assumptions (1.14) and (1.15). There is, however, a gap in Solo's proof, as noted by Zhang [13]. Specifically, Solo's proof made use of the claim (A6) in Appendix I of [8] that for \( \nu \times 1 \) vectors \( x_i \), if \( \sum_{i=1}^{n} x_i^2 = O(n) \) and \( \sum_{i=1}^{n} x_i x_i' \) is nonsingular, then \( x_n' (\sum_{i}^{n} x_i x_i')^{-1} x_n \to 0 \). Zhang [13] found an error in the proof of (A6) and concluded that (A6) is "questionable". In fact, (A6) turns out to be false, as can be seen from the following
example in the scalar case $\nu = 1$. Let $J = \{2, 2^2, 2^3, \ldots \}$ and let $x_n = 1$ if $n \notin J$ and $x_n = n^{1/2}$ if $n \in J$. Then $n \leq \sum_1^n x_i^2 \leq n + \sum_{i:2^i \leq n} 2^i \leq 3n$ for all $n$, and $x_n^2/\sum_1^n x_i^2 \geq 1/3$ for $n \in J$, violating (A6).

For $d$-step ahead prediction, Sin, Goodwin and Bitmead [12] proposed an extension of the AML algorithm to construct adaptive predictors using the following implicit approach. Instead of working with (1.5), they introduced a further reparametrization to facilitate the analysis of the AML algorithm that directly estimates the parameters of this reparametrized model. Applying the division algorithm, they wrote

\begin{equation}
1 = \bar{F}(z)C(z) + z^d \bar{G}(z),
\end{equation}

where $\bar{F}(z) = 1 + \bar{f}_1 z + \cdots + \bar{f}_{d-1} z^{d-1}$ and $\bar{G}(z) = \bar{g}_1 + \bar{g}_2 z + \cdots + \bar{g}_h z^{h-1}$. Let

\begin{equation}
\bar{C}(q^{-1}) = 1 - q^{-d} \bar{G}(q^{-1}) = 1 - \bar{g}_1 q^{-d} - \cdots - \bar{g}_h q^{-d-h+1}.
\end{equation}

Multiplying (1.5) by $\bar{F}(q^{-1})$ gives

\begin{equation}
\bar{C}(q^{-1})\{y_{n+d} - F(q^{-1})e_{n+d}\} = \bar{F}(q^{-1})G(q^{-1})y_n + q^{-(\Delta-d)}F(q^{-1})F(q^{-1})B(q^{-1})u_n,
\end{equation}

in view of (4.4) and (4.5). Therefore, analogous to (1.9), the system (1.1) can be written in the prediction form

\begin{equation}
y_{n+d} = \bar{\theta}' \bar{\psi}_n + \eta_{n+d}, \quad \text{where} \quad \eta_{n+d} = F(q^{-1})e_{n+d} = y_{n+d} - \bar{y}_{n+d},
\end{equation}

\begin{equation}
\bar{\psi}_n = (y_n, \cdots, y_{n-p(d)-d+2}, u_{n-\Delta+d}, \cdots, u_{n-k-\Delta-d+3}, \bar{y}_n, \cdots, \bar{y}_{n-h+1})',
\end{equation}

\begin{equation}
\bar{\theta} = (g_1, g_1 \bar{f}_1 + g_2, \cdots, \bar{f}_{d-1} f_{d-1} b_k, \bar{g}_1, \cdots, \bar{g}_h)'.
\end{equation}

In analogy with (1.13), Sin, Goodwin and Bitmead [12] introduced the following extended least squares algorithm to estimate $\bar{\theta}$:

\begin{align}
&\bar{\theta}_n = \bar{\theta}_{n-d} + P_{n-d} \bar{\phi}_{n-d}(y_n - \bar{\theta}'_{n-d} \bar{\phi}_{n-d}), \\
&P_{n-1} = P_{n-d} - \bar{\phi}_{n-d} \bar{\phi}'_{n-d}, \\
&\bar{\phi}_n = (y_n, \cdots, y_{n-p(d)-d+2}, u_{n-\Delta+d}, \cdots, u_{n-k-\Delta-d+3}, \bar{\theta}'_{n-d} \bar{\phi}_{n-d}, \cdots, \bar{\theta}'_{n-h+1} \bar{\phi}_{n-h+1-d})'.
\end{align}

Thus, (4.8) is an AML-type algorithm which replaces the $\bar{y}_i$ in $\bar{\psi}_n$ by the \textit{a posteriori} $d$-step ahead predictor $\bar{\theta}'_{i} \bar{\phi}_{i-d}$. Note also that (4.8) can be regarded as “interlacing” $d$ unit-delay-type recursions for $\bar{\theta}_{j+d}(j = 1, \cdots, d)$.

The extended least squares $d$-step ahead predictor is

\begin{equation}
\bar{y}_{n+d} = \bar{\theta}'_{n} \bar{\phi}_n,
\end{equation}

21
where $\tilde{\theta}_n$ and $\tilde{\phi}_n$ are given by (4.8). Assuming (1.21) and

$$
(4.10) \quad \tilde{C}(z) \quad \text{is stable and} \quad \min_{|z|=1}(1/\tilde{C}(z) - 1/2) > 0,
$$

Sin, Goodwin and Bitmead [12] modified Solo's [8] argument to prove that $n^{-1} \sum_{i=1}^{n} (\hat{y}_{i+d} - \hat{y}_{i+d})^2 \to 0$ a.s., i.e., the adaptive predictors (4.9) are globally convergent. As noted by Zhang [13], their proof uses Solo's [8] result (A6) to conclude that $\tilde{\phi}_n' P_n \tilde{\phi}_n \to 0$ a.s. Since (A6) has been shown to be invalid, what they have in fact proved is that under the assumptions (4.10) and (1.21),

$$
(4.11) \quad n^{-1} \sum_{i=1}^{n} (\hat{y}_{i+d} - \hat{y}_{i+d})^2 \to 0 \quad \text{a.s. on} \quad \{ \lim_{n \to \infty} \tilde{\phi}_n' P_n \tilde{\phi}_n = 0 \}.
$$

Modifying the proof of Theorem 1 of Lai and Wei [9] (instead of Solo's [8] arguments) with details similar to those provided by [12] leads to the following analog of (4.2), which is considerably stronger than (4.11).

**THEOREM 2.** Suppose that the random disturbances $\epsilon_n$ in the linear stochastic system (1.1) satisfy (1.3) and that the positive real assumption (4.10) holds for the polynomial $\tilde{C}(z)$ defined from $C(z)$ by (4.4) and (4.5). Consider the extended least squares algorithm $\tilde{\theta}_n$, defined by (4.8) for the implicit model (4.7), and its associated adaptive $d$-step ahead predictor (4.9). Then for every $0 < \delta < 1$,

$$
(4.12) \quad \sum_{i=1}^{n} (\hat{y}_{i+d} - \hat{y}_{i+d})^2 I_{\{ \hat{\phi}_i \neq 0 \}} = O(\log \det P_n^{-1}) \quad \text{a.s.},
$$

$$
(4.13) \quad \sum_{i=1}^{n} \| \hat{\phi}_i - \check{\phi}_i \|^2 = O(\log(2 + \sum_{i=1}^{n} \| \hat{\phi}_i \|^2)) \quad \text{a.s.},
$$

where $\check{\phi}_i$ is defined in (4.7). If furthermore (1.21) holds, then $\sum_{i=1}^{n} \hat{\phi}_i^2 = \sum_{i=1}^{n} (y_i - \eta_i)^2 = O(n)$ a.s. and therefore $\sum_{i=1}^{n} \| \hat{\phi}_i \|^2 = O(n)$ a.s. by (4.13), so (4.12) implies that

$$
(4.14) \quad \sum_{i=1}^{n} (\hat{y}_{i+d} - \hat{y}_{i+d})^2 = O(\log n) \quad \text{a.s. on} \quad \{ \lim_{n \to \infty} \tilde{\phi}_n' P_n \tilde{\phi}_n < 1 \}.
$$

Because of the restriction to the event $\{ \lim_{n \to \infty} \tilde{\phi}_n' P_n \tilde{\phi}_n < 1 \}$, (4.11) (or (4.14)) falls short of providing globally convergent adaptive predictors. Zhang [13] suggested the following modification of (4.8a) and (4.8b) to circumvent this difficulty, defining $\tilde{\phi}_n$ as in (4.8c):

$$
(4.15a) \quad \tilde{\theta}_n = \tilde{\theta}_{n-d} + \alpha_n P_{n-d} \tilde{\phi}_{n-d} (y_n - \tilde{\theta}_{n-d} \tilde{\phi}_{n-d}),
$$

$$
(4.15b) \quad r_n = r_{n-1} + \| \tilde{\phi}_n \|^2, \quad R_n = (P_{n-d}^{-1} + \tilde{\phi}_n \tilde{\phi}_n')^{-1},
$$

$$
(4.15c) \quad P_n = R_n \text{ and } \alpha_n = 1, \text{ if } r_n \text{tr}(R_n) \leq K_1 \text{ and } \tilde{\phi}_n' P_{n-d} \tilde{\phi}_n \leq K_2,
$$

$$
P_n = (r_n^{-1} / r_n) P_{n-d} \text{ and } \alpha_n = 1 / (1 + \tilde{\phi}_n' P_{n-d} \tilde{\phi}_n), \text{ otherwise},
$$

22
where $K_1, K_2$ are prescribed constants. Under the assumptions (1.21) and (4.10), he showed that the adaptive predictors $\hat{y}_{n+d} = \hat{\theta}_n^* \hat{\phi}_n$ associated with (4.15) satisfy the global convergence property (1.22). Earlier, in the case $\Delta = d = 1$, Sin and Goodwin [21] introduced a similar recursive identification algorithm, called “modified least squares”, which does not include the condition $\hat{\phi}_n^* P_{n-1} \hat{\phi}_n \leq K_2$ as in (4.15c). They did not need this condition because they applied Solo’s [8] result (A6) to conclude that $\hat{\phi}_n^* P_{n-1} \hat{\phi}_n \to 0$. Zhang [13] found (A6) questionable and further modified the Sin-Goodwin algorithm to patch this gap and to extend to general $d$.

Instead of using the matrix gain (4.8b) in the extended least squares algorithm, the modified least squares algorithm changes this gain by not including $\hat{\phi}_n^* \hat{\phi}_n$ into the sum and deflating the matrix by a scalar multiple whenever certain conditions are not met. Note that the scalar gain $1/r_n = 1/(r_0 + \sum_i^n \|\phi_i\|^2)$ of the stochastic gradient algorithm also appears in the condition $r_n \text{tr}(R_n) \leq K_1$ and in the dampening factor $r_{n-d}/r_n$ of (4.15c).

Our results in Theorem 2 suggest a simpler and more direct way of combining the extended least squares algorithm with the stochastic gradient algorithm to produce a globally convergent adaptive predictor. This is the content of

**COROLLARY 2.** Suppose that the random disturbances $\epsilon_n$ in the linear system (1.1) satisfy (1.3) and that the outputs $y_n$ and inputs $u_n$ satisfy (1.21). Assume that (4.10) holds and that $\min_{|z|=1} \text{Re}\{C(z) - (d-1/2)a\} > 0$ for some $a > 0$. Define the stochastic gradient algorithm $\theta_n$ by (3.12) and the extended least squares algorithm $\hat{\theta}_n$ by (4.8). Let $\hat{y}_{n+d}$ be the adaptive $d$-step ahead predictor associated with the stochastic gradient algorithm, as in (3.12d). Take $0 < \delta < 1$ and define

$$
\hat{y}_{n+d}^* = \hat{\theta}_n^* \hat{\phi}_n \quad \text{if} \quad \hat{\phi}_n^* P_n \hat{\phi}_n \leq \delta,
$$

$$
= \hat{y}_{n+d} \quad \text{if} \quad \hat{\phi}_n^* P_n \hat{\phi}_n > \delta,
$$

where $P_n$ and $\phi_n$ are given in (4.8b) and (4.8c). Then

$$
n^{-1} \sum_{i=1}^n (\tilde{y}_{i+d} - \hat{y}_{i+d}^*)^2 \to 0 \quad \text{a.s.,}
$$

$$
\sum_{i=1}^n I_{\{\hat{\phi}_n^* P_n \hat{\phi}_n > \delta\}} = O(\log n) \quad \text{a.s.}
$$

**Proof.** By Theorem 2, $\sum_1^n \|\phi_i\|^2 = O(n)$ a.s. since (1.21) holds, and

$$
\sum_{i=1}^n (\tilde{y}_{i+d} - \hat{\theta}_n^* \hat{\phi}_n)^2 I_{\{\hat{\phi}_n^* P_n \hat{\phi}_n \leq \delta\}} = O(\log \lambda_{\max}(P_n^{-1})) = O(\log n) \quad \text{a.s.}
$$

By (3.15) of Theorem 1 and (1.21),

$$
\sum_{i=1}^n (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 = o(n) \quad \text{a.s.}
$$

23
From (4.16), (4.19) and (4.20), (4.17) follows.

To prove (4.18), note that for \( \nu = 0, \ldots, d - 1 \),

\[
\sum_{i \leq n, i \equiv \nu \text{ (mod } d)} \tilde{\Phi}_i (P_i^{-1})^{-1} \tilde{\Phi}_i = O\left( \log(2 + \sum_{i=1}^{n} \|\tilde{\Phi}_i\|^2) \right) \quad \text{a.s.,}
\]

by (4.8b) and Lemma 4(ii). Since \( \sum_1^n \|\tilde{\Phi}_i\|^2 = O(n) \) a.s., (4.18) follows from (4.21). \( \blacksquare \)

In view of (4.18), the stochastic gradient component of the adaptive predictor (4.16) is used very infrequently, only at a relative frequency of \( O(n^{-1} \log n) \) within \( n \) stages. At other times the extended least squares component of (4.16) is used and the cumulative squared difference between the adaptive predictor and the optimal predictor at these times is of the order \( O(\log n) \), as in (4.19). Instead of the preceding implicit approach to construct globally convergent adaptive \( d \)-step ahead predictors, one can also use the explicit approach based on the AML algorithm, as in the following.

**COROLLARY 3.** Suppose that the random disturbances \( e_n \) in (1.1) satisfy (1.3) and that the polynomial \( C(z) \) satisfies the positive real condition (1.15). Suppose that the input-output data satisfy (1.21) and the excitation condition

\[
\lambda_{\min}(\Psi_i \Psi_i^t)/\log n \to \infty \quad \text{a.s.,}
\]

where \( \Psi_t = (y_t, \ldots, y_{t-p+1}, u_{t-\Delta+1}, \ldots, u_{t-\Delta-k+2}, e_t, \ldots, e_{t-h+1})' \). Define the AML algorithm \( \Theta_n \) by (1.13) and use it to define the adaptive \( d \)-step ahead predictor \( \hat{y}_{n+d} \) by (1.19). Then \( \{\hat{y}_{n+d}\} \) satisfies the global convergence property (1.22).

**Proof.** In view of (1.21) and (4.22), (1.17) is satisfied and therefore \( \Theta_n \to \Theta \) a.s. by Theorem 1 of [9]. Hence by Lemma 5(ii), (1.20) holds. From (1.20) and (1.21), the desired conclusion (1.22) follows. \( \blacksquare \)

5. **Consistent parameter estimation and monitored recursive maximum likelihood.** In this section we shall assume that \( C(z) \) is stable. Suppose that we are able to find strongly consistent estimators \( \Theta_n \) of the parameter vector \( \Theta \) defined in (1.11). Then under the assumption (1.21), the adaptive \( d \)-step ahead predictor \( \hat{y}_{n+d} \) constructed from \( \Theta_n \) by the explicit approach (1.19) satisfies the global convergence property (1.22). An example of this has been given by Corollary 3 that uses the AML algorithm for consistent estimation of \( \Theta \). Here we shall make use of the consistent estimators \( \Theta_n \) in another way to provide adaptive predictors \( \hat{y}_{n+d} \) that satisfy

\[
\limsup_{n \to \infty} \sum_{i=1}^{n} (\hat{y}_{i+d} - \hat{y}_{i+d})^2 / \log n \leq (2d - 1) \limsup_{i \to \infty} E(\eta_i^2 | \mathcal{F}_{i-d}) \quad \text{a.s.,}
\]

where \( p(d) = p \vee (h - d + 1) \) and \( \eta_i \) is defined in (1.7). Such adaptive predictors are constructed by the implicit approach using the monitored recursive maximum likelihood algorithm defined
below. Note that (5.1) is a much stronger conclusion than the global convergence property (1.22) and that (5.1) represents an extension of (1.33) which corresponds to the case \( d = 1 \) (so that \( \eta_i = \epsilon_i \)) and \( E(\epsilon_i^2 | F_{i-1}) = \sigma^2 \).

The consistent estimators \( \Theta_n \) will be used to provide “confidence sets” \( S_n \) for the parameter vector \( \theta \), defined in (1.8), of the implicit system such that \( S_n \) shrinks to \( \theta \) as \( n \to \infty \). For example, consider the AML algorithm \( \Theta_n \) of Corollary 3 with pseudoregression vectors \( \Phi_n \). Under the assumptions (1.15), (1.21) and (4.22), it follows from Theorem 1 of [9] that

\[
\frac{\log(\sum_1^n \| \Phi_i \|^2)}{\lambda_{\min}(\sum_1^n \Phi_i \Phi_i')} \to 0 \quad \text{a.s.}
\]

and that

\[
\Theta_n - \Theta = O\left( \left\{ \log(\sum_1^n \| \Phi_i \|^2) \right\}^{1/2} / \lambda_{\min}^{1/2}(\sum_1^n \Phi_i \Phi_i') \right) \quad \text{a.s.}
\]

(5.2)

Since the components of \( \theta \) can be expressed as smooth functions of those of \( \Theta \), \( \Theta_n \) induces a strongly consistent estimator \( \hat{\theta}_n \) of \( \theta \); in fact, (5.2) implies that

\[
\hat{\theta}_n - \theta = O\left( \left\{ \log(\sum_1^n \| \Phi_i \|^2) \right\}^{1/2} / \lambda_{\min}^{1/3}(I + \sum_1^n \Phi_i \Phi_i') \right) \quad \text{a.s.}
\]

(5.3)

Hence we can define \( S_n \) to be a cube with center \( \hat{\theta}_n \) and width \( \{ \log(2 + \sum_1^n \| \Phi_i \|^2) \}^{1/3} / \lambda_{\min}^{1/3} \). Then by (5.3),

\[
P\{ \theta \in S_n \} \quad \text{for all large } n = 1,
\]

(5.4)

and as \( n \to \infty \) the width of \( S_n \) converges to 0 a.s.

The consistent estimators \( \hat{\theta}_n \) and the associated confidence sets \( S_n \) need only be updated occasionally at times \( n(1) < n(2) < \cdots \) for monitoring the recursive maximum likelihood algorithm \( \Theta_n \) that we now introduce. The basic ideas underlying the algorithm \( \Theta_n \) are (i) to extend the RML2 algorithm (1.18) to the implicit system (1.9), and (ii) to constrain (monitor) the algorithm so that it lies inside \( S_{n(j)} \) for \( n(j) \leq n < n(j + 1) \). The projection which we use to constrain \( \theta_n \) is taken with respect to the norm induced by the positive definite matrix \( P_{n-d}^{-1} \) defined in (5.6d) below, instead of the usual Euclidean norm. For \( x \in \mathbb{R}^{p(d)+k+d-1+h} \) and \( n(j) \leq n < n(j + 1) \), let \( \pi_n(x) \) denote the unique solution of the quadratic programming problem

\[
(\pi_n(x) - x)' P_{n-d}^{-1}(\pi_n(x) - x) = \min_{y \in S_{n(j)}} \{ (y - x)' P_{n-d}^{-1}(y - x) \},
\]

i.e., \( \pi_n(x) \) is the projection of \( x \) into \( S_{n(j)} \) with respect to the norm induced by \( P_{n-d}^{-1} \). It is convenient to choose \( S_{n(j)} \) to be a cube so that we have linear constraints for the quadratic programming problem (5.5), which can be handled by simple computational methods (cf. [22]).
Define $\theta_n = (\bar{g}_{n,1}, \cdots, \bar{g}_{n,p(d)}, \bar{b}_{n,1}, (\bar{f}b)_{n,2}, \cdots, (\bar{f}b)_{n,k+d-1}, -\bar{c}_{n,1}, \cdots, -\bar{c}_{n,h})' \text{ for } n > n(1) \text{ by the recursion}$

(5.6a) $\theta_n = \pi_n(\theta_{n-1} + P_{n-d}\xi_{n-d}(y_n - \bar{y}_n)),$
(5.6b) $\xi_n + \bar{c}_{n,1}\xi_{n-1} + \cdots + \bar{c}_{n,h}\xi_{n-h} = \phi_n,$ \quad where \quad $\phi_n = (y_n, \cdots, y_{n-p(d)+1}, u_{n-\Delta+d}, \cdots, u_{n-k-\Delta+2}, \bar{y}_{n+d-1}, \cdots, \bar{y}_{n+d-h})'$,
(5.6c) $\bar{y}_{n+d} = \theta_n'\phi_n,$
(5.6d) $P_{n-1}^{-1} = P_{n-1}^{-1} + \xi_n\xi_n' + \rho_n I,$

where

(5.6e) $\rho_n \geq 0 \text{ is } \mathcal{F}_n\text{-measurable with } \sup_n \rho_n < \infty \text{ a.s.}$

The following theorem, which is analogous to Theorem 2 on the extended least squares algorithm, gives basic asymptotic properties of the adaptive predictors (5.6c) associated with the monitored recursive maximum likelihood algorithm. These results are used to establish the conclusion (5.1) for the choice $\rho_n = 1/n$ in (5.6d) under certain conditions on the input-output data and the stopping times $n(j)$ in Corollary 4 below.

THEOREM 3. Suppose that $C(z)$ is stable and that the random disturbances $\varepsilon_n$ in the linear stochastic system (1.1) satisfy assumption (1.3). Let $n(1) < n(2) < \cdots$ be stopping times with respect to $\{\mathcal{F}_i\}$ and let $S_{n(j)}$ be an $\mathcal{F}_{n(j)}\text{-measurable, closed and convex set such that }$

(5.7) $P\{\theta \in S_{n(j)} \text{ for all large } j\} = 1 \text{ and } \lim_{j \to \infty} (\text{diameter of } S_{n(j)}) = 0 \text{ a.s.}$

Define the monitored recursive maximum likelihood algorithm $\theta_n$ by (5.6), where $\pi_n$ is given by (5.5) for $n(j) \leq n < n(j+1)$, and define $\bar{y}_i$ and $\eta_i$ by (1.6) and (1.7).

(i) Suppose that $\sup_n |\varepsilon_n| < \infty \text{ a.s. Then on the event } \{\lambda_{\max}(P_n^{-1}) \to \infty \text{ and } \xi_n'P_n\xi_n \to 0\},$

(5.8) $\sum_{i \leq n} (\bar{y}_{i+d} - \bar{y}_{i+d})^2 \leq (2d - 1) \left\{ \limsup_{i \to \infty} E(\eta_i^2 | \mathcal{F}_{i-d}) + o(1) \right\} \log \det(P_n^{-1})$

$+ o\left( \sum_{i=1}^n \rho_i \right) + \left( \sum_{j:n(j) \leq n+d} \sum_{s=1}^{h+d-1} ||\xi_{n(j)-s}||^2 + ||\xi_{n(j)-s}|| \right) \text{ a.s.}$

(ii) Suppose that the $n(j)$ are stopping times with respect to $\{\mathcal{F}_{i-d+1}\}$ (i.e., $n(j) = t \in \mathcal{F}_{i-d+1}$) and such that $n(j+1) - n(j) \geq d$. Then (5.8) still holds on $\{\lambda_{\max}(P_n^{-1}) \to \infty \text{ and } \xi_n'P_n\xi_n \to 0\}$. 26
Proof. Let \( Q_n = (\theta_n - \theta)' P_{n-d}^{-1}(\theta_n - \theta) \). For \( n(j) \leq n < n(j+1) \), since \( \pi_n(x) \) is the projection of \( x \) into the closed convex set \( S_{n(j)} \) with respect to the norm induced by \( P_{n-d}^{-1} \), it follows from (5.6a) that if \( \theta \in S_{n(j)} \),

\[
Q_n \leq (\theta_{n-1} - \theta + P_{n-d}\xi_{n-d}(y_n - \bar{y}_n))^' P_{n-d}^{-1}(\theta_{n-1} - \theta + P_{n-d}\xi_{n-d}(y_n - \bar{y}_n)) \\
= (\theta_{n-1} - \theta)' P_{n-d}^{-1}(\theta_{n-1} - \theta) + 2\xi_{n-d}^' (\theta_{n-1} - \theta)(y_n - \bar{y}_n) + \xi_{n-d}' P_{n-d}\xi_{n-d}(y_n - \bar{y}_n)^2 \\
= Q_{n-1} + \rho_{n-d}(\theta_{n-1} - \theta)^2 + [\xi_{n-d}' (\theta_{n-1} - \theta)]^2 + 2\xi_{n-d}' (\theta_{n-1} - \theta)(y_n - \bar{y}_n) \\
+ \xi_{n-d}' P_{n-d}\xi_{n-d}(y_n - \bar{y}_n)^2,
\]

noting that \( P_{i-1}^{-1} = P_{i-1}^{-1} + \xi_i \xi_i' + \rho_i I \). Therefore on the event \( E \triangleq \{ \theta \not\in S_{n(j)} \} \) for finitely many \( j \)'s,

\[
Q_n \leq \sum_{i=n(1)+1}^n [\xi_{i-d}' (\theta_{i-1} - \theta)]^2 + 2 \sum_{i=n(1)+1}^n \xi_{i-d}' (\theta_{i-1} - \theta)(y_i - \bar{y}_i) \\
+ \sum_{i=n(1)+1}^n \xi_{i-d}' P_{i-d}\xi_{i-d}(y_i - \bar{y}_i)^2 + \sum_{i=n(1)+1}^n \rho_{i-d}(\theta_{i-1} - \theta)^2 + O(1) \\
\leq - \sum_{i=n(1)+1}^n [\xi_{i-d}' (\theta_{i-1} - \theta)]^2 + 2 \sum_{i=n(1)+1}^n \xi_{i-d}' (\theta_{i-1} - \theta)[\xi_{i-d}' (\theta_{i-1} - \theta) + (y_i - \bar{y}_i - \eta_i)] \\
+ 2 \sum_{i=n(1)+1}^n \xi_{i-d}' (\theta_{i-1} - \theta)\eta_i + \sum_{i=n(1)+1}^n \xi_{i-d}' P_{i-d}\xi_{i-d}(y_i - \bar{y}_i)^2 + o(\sum_{i=1}^{n-d}) + O(1) \text{ a.s.,}
\]

noting that \( \theta_{i-1} - \theta \to 0 \text{ a.s. by (5.7).} \)

Let \( C_t(q^{-1}) = 1 + \hat{c}_{t,1}(q^{-1}) + \cdots + \hat{c}_{t,h} q^{-h} \), \( \hat{c}_{t,0} = 1 \). From (2.3),

\[
C(q^{-1})[\xi_{i-d}' (\theta_{i-1} - \theta) + (y_i - \bar{y}_i - \eta_i)] = C(q^{-1})[\xi_{i-d}' (\theta_{i-1} - \theta)] - \phi_{i-d}' (\theta_{i-d} - \theta) \\
= [C(q^{-1}) - C_{i-d}(q^{-1})][\xi_{i-d}' (\theta_{i-1} - \theta)] + \sum_{r=1}^h \hat{c}_{i-d,r} \xi_{i-d-r}(\theta_{i-1-r} - \theta_{i-1}) \\
+ \phi_{i-d}' (\theta_{i-1} - \theta_{i-d}),
\]

noting that \( C_{i-d}(q^{-1}) \xi_{i-d} = \phi_{i-d}' \theta \) and \( (\sum_{r=0}^h \hat{c}_{i-d,r} \xi_{i-d-r}) \theta_{i-1} = \phi_{i-d}' \theta_{i-1} \) by (5.6b). Since \( C(z) \) is stable and \( \hat{c}_{i-d,r} \to c_r \text{ a.s. as } i \to \infty \text{ for } r = 1, \cdots, h \), and since \( \theta_{i-1} - \theta_{i-s} = \sum_{j=i-s+1}^{i-1} (\theta_j - \theta_{j-1}) \text{ for } s \geq 1 \), it follows from (5.11), (5.6b) and Lemma 5(i) that

\[
\sum_{i=n(1)+1}^n [\xi_{i-d}' (\theta_{i-1} - \theta) + (y_i - \bar{y}_i - \eta_i)]^2 = o\left( \sum_{i=n(1)+1}^n [\xi_{i-d}' (\theta_{i-1} - \theta)]^2 \right) + O(1) \\
+ O(\sum_{i=1}^{n} \sum_{r=-d+1}^{h-1} [\xi_{i-d-r}(\theta_i - \theta_{i-1})]^2) \text{ a.s.}
\]
For $n(j) < i < n(j + 1)$, $\theta_{i-1} \in S_{n(j)}$ and therefore $(\theta_i - \theta_{i-1})' P_{i-d}^{-1}(\theta_i - \theta_{i-1}) \leq \xi_{i-d}' P_{i-d} \xi_{i-d} (y_i - \tilde{y}_i)^2$ by (5.6a). Hence by the Schwarz inequality, on $E_1 \triangleq E \cap \{\lim_{n \to \infty} \xi_n' P_n \xi_n = 0\}$,

\begin{equation}
\sum_{i=n(1)+1}^{n} \sum_{i=1}^{h-1} \left[ \xi_{i-d-r}' P_{i-d}^{-1/2} \xi_{i-d-r} (\theta_i - \theta_{i-1}) \right]^2 \leq \sum_{i=n(1)+1}^{n} \left( \sum_{r=-d+1}^{h-1} \xi_{i-d-r}' P_{i-d} \xi_{i-d-r} \right) \xi_{i-d}' P_{i-d} \xi_{i-d} (y_i - \tilde{y}_i)^2
= o\left( \sum_{i=n(1)+1}^{n} \xi_{i-d}' P_{i-d} \xi_{i-d} (y_i - \tilde{y}_i)^2 \right) + O(1)
\end{equation}

by Lemma 4(ii). Hence by (5.12) and (5.13),

\begin{equation}
\sum_{i=n(1)+1}^{n} \left[ \xi_{i-d}' (\theta_{i-1} - \theta) + (y_i - \tilde{y}_i - \eta_i) \right]^2 = o\left( \sum_{i=n(1)+1}^{n} \left[ \xi_{i-d}' (\theta_{i-1} - \theta) \right]^2 \right) + O(1)
+ o\left( \sum_{i=n(1)+1}^{n} \xi_{i-d}' P_{i-d} \xi_{i-d} (y_i - \tilde{y}_i)^2 \right) + o\left( \sum_{j:n(j) \leq n}^{h-1} \sum_{r=-d+1}^{h-1} \|\xi_{n(j)-d-r}\|^2 \right) \text{ a.s. on } E_1,
\end{equation}

noting that $\theta_j - \theta_{j-1} \to 0$ a.s. in view of (5.7).

Since $(y_i - \tilde{y}_i)^2 = (y_i - \tilde{y}_i - \eta_i)^2 + 2(y_i - \tilde{y}_i - \eta_i)\eta_i + \eta_i^2$ and since $y_i - \tilde{y}_i - \eta_i$ and $\xi_{i-d}' P_{i-d} \xi_{i-d} \leq 1$ are $F_{i-d}$-measurable, it follows from Lemma 1 that on $E_1$,

\begin{equation}
\sum_{i=n(1)+1}^{n} \xi_{i-d}' P_{i-d} \xi_{i-d} (y_i - \tilde{y}_i)^2 \leq \left\{ \limsup_{i \to \infty} E(\eta_i^2 | F_{i-d}) + o(1) \right\} \sum_{i=n(1)+1}^{n} \xi_{i-d}' P_{i-d} \xi_{i-d} + o\left( \sum_{i=n(1)+1}^{n} (y_i - \tilde{y}_i - \eta_i)^2 \right) + O(1)
= \left\{ \limsup_{i \to \infty} E(\eta_i^2 | F_{i-d}) + o(1) \right\} \log \det P_{n-d}^{-1} + o\left( \sum_{i=n(1)+1}^{n} (y_i - \tilde{y}_i - \eta_i)^2 \right) + O(1) \text{ a.s.},
\end{equation}

by Lemma 4(i). Let $0 < \lambda < 1/2$. Using the inequality $A^2 \leq (1 + \lambda^2)B^2 + (1 + \lambda^{-2})(A + B)^2$, we obtain from (5.14) and (5.15) that on $E_1$,

\begin{equation}
\sum_{i=n(1)+1}^{n} (y_i - \tilde{y}_i - \eta_i)^2 \leq (1 + \lambda^2 + o(1)) \sum_{i=n(1)+1}^{n} \left[ \xi_{i-d}' (\theta_{i-1} - \theta) \right]^2
+ o(\log \det P_{n-d}^{-1}) + o\left( \sum_{j:n(j) \leq n}^{h+d-1} \sum_{s=1}^{h+d-1} \|\xi_{n(j)-s}\|^2 \right) + O(1) \text{ a.s.}
\end{equation}
Moreover, using the inequality $AB \leq (\lambda^2 A^2 + \lambda^{-2} B^2)/2$, we obtain from (5.14) and (5.15) that on $E_1$

$$
(5.17) \quad \sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-1} - \theta)][\xi'_{i-d}(\theta_{i-1} - \theta) + (y_i - \tilde{y}_i - \eta_i)]
$$
$$\leq (\lambda^2 + o(1)) \sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-1} - \theta)]^2
$$
$$+ o(\log \det \mathbb{P}_{n-d}^{-1}) + o(\sum_{j:n(j) \leq n} \sum_{s=1}^{h+d-1} \|\xi_{n(j)-s}\|^2) + o(\sum_{i=1}^{n-d} \rho_i) + O(1) \quad \text{a.s.}
$$

Writing $\theta_{i-1} - \theta = \theta_{i-d} - \theta + \sum_{r=1}^{d-1}(\theta_{i-r} - \theta_{i-r-1})$, we now proceed to show that on $E_1$,

$$
(5.18) \quad 2| \sum_{i=n(1)+1}^{n} \xi'_{i-d}(\theta_{i-1} - \theta)\eta_{i}| \leq 2(d-1) \{\limsup_{i \to \infty} E(\eta_i^2 | \mathcal{F}_{i-d}) + o(1)\} \log \det \mathbb{P}_{n-d}^{-1}
$$
$$+ o(\sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-1} - \theta)]) + o(\sum_{j:n(j) \leq n} \sum_{s=1}^{h+d-1} \|\xi_{n(j)-s}\|^2)
$$
$$+ o(\sum_{j:n(j) \leq n} \sum_{r=1}^{d-1} \|\xi_{n(j)+r-d}\|\|\eta_{n(j)+r}\|) + O(1) \quad \text{a.s.}
$$

Since $\xi'_{i-d}(\theta_{i-d} - \theta)$ is $\mathcal{F}_{i-d}$-measurable and since $\eta_i = \epsilon_i + f_1 \epsilon_{i-1} + \cdots + f_{d-1} \epsilon_{i-d+1}$, an application of Lemma 1(i) gives

$$
(5.19) \quad \sum_{i=n(1)+1}^{n} \xi'_{i-d}(\theta_{i-d} - \theta)\eta_{i} = O(\sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-d} - \theta)]^2) + O(1)
$$
$$= o(\sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-1} - \theta)]^2) + o(\sum_{r=1}^{d-1} \sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-r} - \theta_{i-r-1})]^2) + O(1) \quad \text{a.s.}
$$

For fixed $r = 1, \cdots, d-1$, we have analogous to (5.13) that on $E_1$

$$
(5.20) \quad \sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-r} - \theta_{i-r-1})]^2 = o(\sum_{i=n(1)+1}^{n-r} \xi'_{i-d}P_{i-d} \xi_{i-d}(y_i - \tilde{y}_i)^2)
$$
$$+ o(\sum_{j:n(j) \leq n-r} \|\xi_{n(j)+r-d}\|^2) + O(1).
$$
Moreover, since \( \theta_{t-1} \in S_{n(j)} \) if \( n(j) < t < n(j+1) \), we obtain by the Schwarz inequality and (5.6a) that

\[
(5.21) \quad \sum_{i \in \{n(1)+r,n(2)+r,\ldots\}} |\xi_{i-d}^{1/2} P_{i-r-d}^{1/2} P_{i-r-d}^{-1/2} (\theta_{i-r} - \theta_{i-r-1}) \eta_i| \leq \sum_{i \in \{n(1)+r,n(2)+r,\ldots\}} \left( \xi_{i-d}^{1/2} P_{i-r-d} \xi_{i-d} \right)^{1/2} \left( \xi_{i-r-d}^{1/2} P_{i-r-d} \xi_{i-r-d} \right)^{1/2} |\eta_i| ||y_{i-r} - \hat{y}_{i-r}|| \leq \sum_{i \leq n} \xi_{i-d}^{1/2} P_{i-r-d} \xi_{i-d} \eta_i^2 / 2 + \sum_{i \leq n} \xi_{i-r-d}^{1/2} P_{i-r-d} \xi_{i-r-d} (y_{i-r} - \hat{y}_{i-r})^2 / 2.
\]

Applying Lemma 1(ii), Lemma 4(ii) and (5.15) to (5.20), and combining the result with (5.19), (5.20), (5.15) and (5.16), we obtain (5.18).

Suppose that \( \sup_n |\varepsilon_n| < \infty \) a.s. Then \( \sup_n |\eta_n| < \infty \) a.s. and therefore

\[
(5.22) \quad \sum_{j:n(j) \leq n} \sum_{r=1}^{d-1} ||\xi_{n(j)+r-d}|| |\eta_{n(j)+r}| = O\left( \sum_{j:n(j) \leq n} \sum_{s=1}^{d-1} ||\xi_{n(j)-s}|| \right) \text{ a.s.}
\]

Now assume that the \( n(j) \) are stopping times with respect to \( \{\mathcal{F}_{i-d+1}\} \). Then for \( r = 1, \ldots, d-1 \), \( T_j \triangleq n(j) + r - d \) is a stopping time with respect to \( \{\mathcal{F}_i\} \) and \( \eta_{n(j)+r} = \sum_{i=1}^{d} f_{d-i} \varepsilon_{T_j+i} + f_0 = 0 \). Moreover, \( ||\xi_{n(j)+r-d}|| \) is \( \mathcal{F}_{T_j} \)-measurable. Hence an application of Lemma 1(iii) and (i) gives

\[
(5.22') \quad \sum_{j:n(j) \leq n} \sum_{r=1}^{d-1} ||\xi_{n(j)+r-d}|| |\eta_{n(j)+r}| = \sum_{j:n(j) \leq n} \sum_{r=1}^{d-1} ||\xi_{n(j)+r-d}|| E\left( \sum_{i=1}^{d} f_{d-i} \varepsilon_{T_j+i} \right | \mathcal{F}_{T_j})
\]

\[ + o\left( \sum_{j:n(j) \leq n} \sum_{r=1}^{d-1} ||\xi_{n(j)+r-d}||^2 \right) + O(1) \text{ a.s.}
\]

Since \( E\left( \sum_{i=1}^{d} f_{d-i} \varepsilon_{T_j+i} \right | \mathcal{F}_{T_j}) \leq E^{1/\alpha}\left( \sum_{i=1}^{d} f_{d-i} \varepsilon_{T_j+i} \right | \mathcal{F}_{T_j}) \), either (5.22') or (5.22) implies that

\[
(5.23) \quad \sum_{j:n(j) \leq n} \sum_{r=1}^{d-1} ||\xi_{n(j)+r-d}|| |\eta_{n(j)+r}| = O\left( \sum_{j:n(j) \leq n} \sum_{s=1}^{d-1} \left( ||\xi_{n(j)-s}|| + ||\xi_{n(j)-s}||^2 \right) \right) + O(1) \text{ a.s.}
\]

Applying (5.15)–(5.18) and (5.23) to (5.10), we obtain that on \( \mathcal{E}_1 \),

\[
Q_n + (1 - 2\lambda^2 + o(1)) \sum_{i=n(1)+1}^{n} [\xi_{i-d}(\theta_{i-1} - \theta)]^2
\]

\[
\leq (2d - 1) \{ \limsup_{i \to \infty} E(\eta_i^2 | T_{i-d}^n) + o(1) \} \log \det P_{n-d}^{-1}
\]

\[ + o\left( \sum_{j:n(j) \leq n} \sum_{r=1}^{h+d-1} \left( ||\xi_{n(j)-r}|| + ||\xi_{n(j)-r}||^2 \right) \right) + O(1) \text{ a.s.}
\]
Since $Q_n \geq 0$, it follows from (5.24) and (5.16) that on $E_1$

\begin{equation}
(5.25) \quad \sum_{i=n(1)}^{n} (\tilde{y}_i - \tilde{y}_i)^2 = \sum_{i=n(1)}^{n} (\tilde{y}_i - \tilde{y}_i - \eta_i)^2 \\
\leq (1 + \lambda^2 + o(1))(1 - 2\lambda^2 + o(1))^{-1} (2d - 1) \limsup_{i \to \infty} E(\eta_i^2 | \mathcal{F}_{i-d}) + o(1) \log \det P_{n-d}^{-1} \\
+ o(\sum_{j:n(j) \leq n} \sum_{r=1}^{h+d-1} (\|\xi_{n(j)-r}\| + \|\xi_{n(j)-r}\|^2)) + o(\sum_{i=1}^{n-d} \rho_i) + O(1) \quad \text{a.s.}
\end{equation}

Since $\lambda$ can be arbitrarily small, (5.25) implies that (5.8) holds on $E_1 \cap \{\log \det P_n^{-1} \to \infty\} = E \cap \{\xi_n' P_n \xi_n \to 0 \text{ and } \lambda_{\max}(P_n^{-1}) \to \infty\}$. Since $P(E) = 1$ by (5.7), the desired conclusion follows. ■

**COROLLARY 4.** With the same notation and assumptions as in Theorem 3(i), suppose that the input-output data satisfy

\begin{equation}
(5.26) \quad \sum_{i=1}^{n}(y_i^2 + u_i^2) = O(n) \quad \text{and} \quad \max_{i \leq n}(y_i^2 + u_i^2) = o(\log n) \quad \text{a.s.}
\end{equation}

Then

\begin{equation}
(5.27) \quad \sum_{i=1}^{n}\|\xi_i\|^2 = O(n) \quad \text{and} \quad \|\xi_n\|^2 = o(\log n) \quad \text{a.s.}
\end{equation}

Suppose furthermore that $\rho_n = 1/n$ and that the stopping times $n(1) < n(2) < \cdots$ are so chosen that

\begin{equation}
(5.28) \quad \sum_{j:n(j) \leq n} \max_{1 \leq r \leq h+d-1} (\|\xi_{n(j)-r}\| + \|\xi_{n(j)-r}\|^2) = O(\log n) \quad \text{a.s.,}
\end{equation}

(which is possible in view of (5.27)). Then (5.1) holds.

**Proof.** Since $\tilde{y}_i = y_i - \eta_i$ and since $\sup_i |\epsilon_i| < \infty$ a.s., it follows from (5.26) that

\begin{equation}
(5.29) \quad \sum_{i=1}^{n}\|\psi_i\|^2 = O(n) \quad \text{and} \quad \max_{i \leq n}\|\psi_i\|^2 = o(\log n) \quad \text{a.s.},
\end{equation}

where $\psi_i$ is defined in (1.9). By Lemma 2,

\begin{equation}
(5.30) \quad C(q^{-1})(y_n - \tilde{y}_n - \eta_n) = -(\theta_{n-d} - \theta)'(\phi_{n-d} - \psi_{n-d}) - (\theta_{n-d} - \theta)'\psi_{n-d}, \quad \text{and therefore}
\end{equation}

\begin{equation}
(5.31) \quad \tilde{C}_{n-d}(q^{-1})(y_n - \tilde{y}_n - \eta_n) = -(\theta_{n-d} - \theta)'\psi_{n-d},
\end{equation}

31
noting that

$$\phi_{n-d} - \psi_{n-d} = -(0, \cdots, 0, y_{n-1} - \hat{y}_{n-1} - \eta_{n-1}, \cdots, y_{n-h} - \hat{y}_{n-h} - \eta_{n-h})'$$

Since $\theta_{n-d} \to \theta$ a.s., it follows from (5.29), (5.30) and Lemma 5(i) that

$$\sum_{i=1}^{n} (y_i - \hat{y}_i - \eta_i)^2 = O\left(\sum_{i=1}^{n} \|\theta_{i-d} - \theta\|^2 \|\psi_{i-d}\|^2\right) = o(n) \quad \text{a.s.,}$$

$$y_n - \hat{y}_n - \eta_n = O\left(\max_{i \leq n} \|\theta_{i-d} - \theta\| \|\psi_{i-d}\|\right) = o((\log n)^{1/2}) \quad \text{a.s.}$$

In view of (5.31), it follows from (5.29), (5.32) and (5.33) that

$$\sum_{i=1}^{n} \|\phi_i\|^2 = O(n) \quad \text{and} \quad \max_{i \leq n} \|\phi_i\|^2 = o(\log n) \quad \text{a.s.}$$

Since $\hat{C}_n(q^{-1})\xi_n = \phi_n$ by (5.6b), (5.27) follows from (5.34) and Lemma 5(ii).

Suppose that $\rho_n = 1/n$ and that (5.28) holds. Since $P_n^{-1} = P_{n(1)}^{-1} + \sum_{j=n(1)+1}^{n} \xi_j \xi_j' + \sum_{j=n(1)+1}^{n} I/j, \lambda_{\min}(P_n^{-1}) \geq (1 + o(1)) \log n \to \infty$ and $\xi_n P_n \xi_n \leq \|\xi_n\|^2/\lambda_{\min}(P_n^{-1}) \to 0$ a.s. by (5.27). Since $\log \det P_n^{-1} \leq (p(d) + k + d - 1 + h) \log \lambda_{\max}(P_n^{-1})$, the desired conclusion (5.1) follows from (5.8), (5.27) and (5.28). \mm

In the preceding implicit (or direct) approach to adaptive prediction using the monitored recursive maximum likelihood algorithm, we have introduced the summand $\rho_n I$ to form $P_n^{-1}$ in (5.6d) in order to get around the difficulties with over-parametrization inherent in the implicit approach. Corollary 4 shows the usefulness of this idea in resolving such difficulties. To avoid over-parametrization, it is also natural to try the explicit or indirect approach instead, involving the monitored recursive maximum likelihood estimator $\Theta_n$ of the parameter vector $\Theta = (-a_1, \cdots, -a_p, b_1, \cdots, b_k, c_1, \cdots, c_h)'$ for the system (1.1):

$$\Theta_n = \pi_n(\Theta_{n-1} + P_{n-1} \zeta_{n-1} \epsilon_n),$$

$$\zeta_n + \hat{c}_n, \epsilon_{n-1} + \cdots + \hat{c}_n \zeta_{n-h} = \Phi_n,$$  \quad where

$$\Phi_n = (y_n, \cdots, y_{n-p+1}, u_{n-\Delta+1}, \cdots, u_{n-\Delta+k+2}, \epsilon_n, \cdots, \epsilon_{n-h+1})'$$

$$\epsilon_n = y_n - \Theta_n^{-1} \Phi_n,$$

$$P_n^{-1} = P_{n-1}^{-1} + \zeta_n \zeta_n'$$

The $\pi_n$ in (5.35a) is the projection, with respect to the norm induced by $P_{n-1}^{-1} = P_0^{-1} + \sum_{i=1}^{n-1} \zeta_i \zeta_i'$, into a closed convex subset $S_n(j)$ of $R^{p+k+h}$. An analog of Theorem 3 for the adaptive 1-step ahead predictors $\hat{y}_{n+1} = \Theta_n' \Phi_n$ can be proved by similar (and simpler) arguments and is given in the following.
THEOREM 4. Suppose that $C(z)$ is stable and that the random disturbances $\epsilon_n$ in the linear stochastic system (1.1) satisfy assumption (1.3). Let $n(1) < n(2) < \cdots$ be stopping times with respect to $\{F_t\}$ and let $S_{n(j)}$ be an $F_{n(j)}$-measurable, closed and convex subset of $R^{p+k+h}$ such that (5.7) holds. Define the monitored recursive maximum likelihood estimator $\Theta_n$ by (5.35), where $\pi_n$ is given by (5.5) (with $d = 1$ and $S_{n(j)} \subset R^{p+k+h}$) for $n(j) \leq n < n(j + 1)$, and define $\tilde{\Theta}$ and $\Psi_t$ by (1.11) and (1.12). Then on $\{\lambda_{\max}(P^{-1}_n) \rightarrow \infty$ and $\zeta_n'P_n\zeta_n \rightarrow 0\}$,

$$
\sum_{i \leq n}(\Theta_i'Y_i - \Theta_i'\Phi_i)^2 \leq \limsup_{i \rightarrow \infty} E\{\epsilon_i^2|F_{i-1}\} + o(1) \log \det P^{-1}_n
$$

$$
+ o\left(\sum_{j:n(j) \leq n+1} \sum_{r=1}^h \|\xi_{n(j)-r}\|^2\right) \quad \text{a.s.}
$$

For $d > 1$, analysis of the adaptive $d$-step ahead predictors defined from $\Theta_n$ using the indirect approach (1.19) becomes much more difficult because of the inherent nonlinearities in (1.19) and (1.4), as illustrated by the following example. Consider 2-step ahead prediction in the ARMAX model

$$
y_n = \alpha y_{n-1} + \beta u_{n-2} + \epsilon_n + \epsilon_{n-1}.
$$

Here $C(z) = 1+cz$, $A(z) = 1-az$, and therefore by (1.4), $F(z) = 1+(a+c)z$, $G(z) = a^2+ac$, so

$$
\tilde{y}_{n+2} + c\tilde{y}_{n+1} = (\alpha^2 + ac)y_n + \beta u_n + \beta(\alpha + c)u_{n-1}.
$$

Since $\Theta_n = (\tilde{\alpha}_n, \tilde{\beta}_n, \tilde{\epsilon}_n)' \rightarrow \Theta$ a.s. by (5.7),

$$
\tilde{\alpha}_n^2 - a^2 \sim 2\alpha(\tilde{\alpha}_n - \alpha), \quad \tilde{\alpha}_n\tilde{\epsilon}_n - ac \sim c(\tilde{\alpha}_n - \alpha) + \alpha(\tilde{\epsilon}_n - c), \quad \text{etc.}
$$

Applying the Taylor expansions (5.39) to (5.38) and to a similar expression for $\tilde{y}_{n+2}$ gives

$$
\tilde{y}_{n+2} - \tilde{y}_{n+2} + \tilde{\epsilon}_n(\tilde{y}_{n+1} - \tilde{y}_{n+1}) \sim (\tilde{\alpha}_n - \alpha)\{(2\alpha + c)y_n + \beta u_{n-1}\}
$$

$$
+ (\tilde{\beta}_n - \beta)\{u_n + (\alpha + c)u_{n-1}\} + (\tilde{\epsilon}_n - c)(\alpha y_n + \beta u_{n-1} - \tilde{y}_{n+1}).
$$

From (5.38), it follows that

$$
C(q^{-1})(\tilde{y}_{n+1} - \alpha y_n - \beta u_{n-1}) = \alpha^2 y_{n-1} + \beta\alpha u_{n-2} - \alpha y_n
$$

$$
- \alpha(y_n - \alpha y_{n-1} - \beta u_{n-2}) = -\alpha[C(q^{-1})\epsilon_n], \quad \text{by (5.37)},
$$

and therefore $\alpha y_n + \beta u_{n-1} - \tilde{y}_{n+1} = \alpha \epsilon_n$. Noting that $\Psi_n = (y_n, u_{n-1}, \epsilon_n)'$ and that $\alpha y_n + \beta u_{n-1} = y_{n+1} - (\epsilon_{n+1} + \epsilon n$ by (5.37), it follows from (5.40) that

$$
(1 + \tilde{\epsilon}_n q^{-1})(\tilde{y}_{n+2} - \tilde{y}_{n+2}) \sim \alpha(\Theta_n - \Theta)'\Psi_n + (\tilde{\beta}_n - \beta')(u_n + c u_{n-1})
$$

$$
+ (\tilde{\alpha}_n - \alpha)(y_{n+1} + c y_n - \epsilon_{n+1} - \epsilon n).
$$

33
Since \( y_{n+1} - \epsilon_{n+1} = \Theta' \Psi_n \), we have analogous to (5.31) that
\[
(\Theta_n - \Theta)' \Psi_n = -(1 + \hat{c}_n q^{-1}) (\Theta' \Psi_n - \Theta'_n \Phi_n),
\]
where \( \Phi_n = (y_n, u_{n-1}, \epsilon_n)' \) and \( \epsilon_n = y_n - \Theta'_{n-1} \Phi_{n-1} \). Moreover,
\[
\begin{align*}
(\hat{\alpha}_n - \alpha)(y_{n+1} + cy_n - \epsilon_{n+1} - c\epsilon_n) + (\hat{\beta}_n - \beta)(u_n + cu_{n-1}) \\
= C(q^{-1}) \{(\Theta_n - \Theta)' \Psi_n\} - (\hat{\alpha}_n - \hat{\alpha}_n, \hat{\beta}_n - \hat{\beta}_n, c_{n+1} - \hat{c}_n) \Psi_{n+1} \\
- C(q^{-1}) \{(\hat{\alpha}_n - \alpha + \hat{c}_n - c)\epsilon_{n+1}\} - c(\hat{\alpha}_n - \alpha_{n-1} + \hat{c}_n - \hat{c}_{n-1})\epsilon_n.
\end{align*}
\]

In view of (5.41)-(5.43) together with Lemmas 5(i) and 1(ii), we have therefore reduced the analysis of \( \sum_1^N (\tilde{y}_{n+2} - \tilde{y}_{n+2})^2 \) to that of \( \sum_1^{N+1} (\Theta' \Psi_n - \Theta'_n \Phi_n)^2 \), which has been studied in Theorem 4, plus the additional terms \( \sum_1^{N+1} (\Theta_n - \Theta_{n-1})' \Psi_n^2 \), \( \sum_1^N (\hat{\alpha}_n - \hat{\alpha}_n - \hat{c}_n - c)^2 \) and \( \sum_1^N (\hat{\alpha}_n - \hat{\alpha}_n - \hat{c}_n - \hat{c}_n)^2 \epsilon_n^2 \). However, for general polynomials \( A(z) \), \( B(z) \) and \( C(z) \) and general \( d \geq 2 \), the algebra involved in a similar reduction becomes unwieldy and leads to more complicated additional terms, and it is doubtful that the indirect approach would provide sharp results of the kind in Corollary 4.

6. Multivariable extensions and control applications. The recursive estimators \( \Theta_n \) of the parameter vector \( \Theta \) of the linear stochastic system (1.1) in §§3-5 can obviously be extended to the multivariable setting in which the \( y_n, \epsilon_n \) and \( u_n \) in (1.1) are \( \nu \times 1 \) vectors, and \( A(q^{-1}) = I + A_1 q^{-1} + \cdots + A_p q^{-p} \), \( B(q^{-1}) = B_1 + \cdots + B_k q^{-k-1} \) and \( C(q^{-1}) = I + C_1 q^{-1} + \cdots + C_h q^{-h} \) are \( \nu \times \nu \) matrix polynomials in the backward shift operator \( q^{-1} \).

Letting \( \Theta = (-A_1, \cdots, -A_p, B_1, \cdots, B_k, C_1, \cdots, C_h)' \), we can again write (1.1) in the regression form \( y_n = \Theta' \Psi_{n-1} + \epsilon_n \), where we now define
\[
\Psi_n = (y'_{n}, \cdots, y'_{n-p+1}, u'_{n-\Delta+1}, \cdots, u'_{n-\Delta-k+2}, \epsilon'_{n}, \cdots, \epsilon'_{n-h+1}),
\]
\[
\Phi_n = (y'_{n}, \cdots, y'_{n-p+1}, u'_{n-\Delta+1}, \cdots, u'_{n-\Delta-k+2}, \epsilon'_{n}, \cdots, \epsilon'_{n-h+1}).
\]

With \( \Phi_n \) defined by (6.2), we define the stochastic gradient algorithm by (1.30). Then the conclusions of Corollary 1 still hold under the assumption
\[
C(e^{it}) + C'(e^{-it}) - aI \quad \text{is positive definite for all } |t| \leq \pi.
\]

Note that in the scalar case \( \nu = 1 \), (6.3) reduces to the assumption \( \min_{|z|=1} \text{Re} \{C(z) - a/2\} > 0 \) of Corollary 1, and that Lemma 3, which is a key tool in the proof of Corollary 1, has in fact been stated for the multivariate positive real assumption (6.3).

With \( \Phi_n \) defined by (6.2) and \( \epsilon_t = y_t - \Theta'_t \Phi_{t-1} \), define the AML algorithm \( \Theta_n \) by (1.13a) and replace the positive real condition (1.15) by its multivariable version
\[
\det (C(z)) \neq 0 \quad \text{for } |z| \leq 1 \quad \text{(i.e. } C(z) \text{ is stable), and}
\]
\[
C^{-1}(z) + (C^{-1}(z))^* - I \quad \text{is positive definite for all } |z| = 1,
\]

34
where * denotes conjugate transpose. Then as shown by Lai and Wei [23], \( \Theta_n \to \Theta \) a.s. and (4.1) holds under (6.4) and (1.17). Note that the extended stochastic Lyapunov function (1.16) now takes the form \( Q_n = \text{tr}\left\{ (\Theta_n - \Theta) P_n^{-1} (\Theta_n - \Theta)' \right\} \), as in [23].

With \( \Phi_n \) defined by (6.2) and \( S_n(j) \subset R^{n(p+k+h)} \), define the monitored recursive maximum likelihood algorithm \( \Theta_n \) by (5.35). Then Theorem 4 on the adaptive 1-step ahead predictors \( \hat{y}_{n+1} = \Theta'_n \Phi_n \) still holds, with \( \| \hat{y}_{i+1} - \Theta'_i \hat{y}_i \|^2 \) in place of \( (\Theta'_i \hat{y}_i - \Theta'_i \hat{y}_i)^2 \) and with \( E(\| \epsilon_i \|^2 | F_{i-1}) \) in place of \( E(\epsilon_i^2 | F_{i-1}) \) in (5.36).

For \( d \geq 1 \), the minimum variance \( d \)-step ahead predictors \( \tilde{y}_{n+d} \) in multivariable systems are considerably more complicated than (1.6) because of the non-commutativity of the matrix polynomials, cf. [3, pp. 270-271]. However, in the special case \( C_i = c_i I \) \((i = 1, \ldots, h)\), for which \( C(q^{-1}) \) acts like a scalar polynomial, the recursive representation (1.6) for \( \tilde{y}_{n+d} \) still holds. Letting \( y_n(1), \ldots, y_n(\nu) \) denote the components of \( y_n \), we can write (1.6) and (1.7) as

\[
(6.5) \quad y_{n+d}(r) = \psi_{n,r}(r) + \eta_{n+d}(r), \quad r = 1, \ldots, \nu,
\]

where

\[
\psi_{n,r} = (y_n', \ldots, y'_n, u_n', \ldots, u_n', \ldots, y_n', \ldots, y_n', \ldots, y_n', \ldots, \hat{y}_{n+d}(r)),
\]

\( \tilde{y}_i(r) \) and \( \eta_i(r) \) are the components of \( \tilde{y}_i \) and \( \eta_i \), respectively, and \( \theta(r) \) has the same form as \( \theta \) in (1.8) but with \( g_i \) or \( (fb)_i \) replaced by the \( r \)-th row of the matrix \( G_i \) or \( (FB)_i \). Therefore we can apply the monitored recursive maximum likelihood algorithm \( \theta_n(r) \) as defined in (5.6) to estimate \( \theta(r) \) and thereby construct the adaptive predictor \( \tilde{y}_{n+d}(r) = \psi_{n,r}(r) \) of \( y_{n+d}(r) \).

By Theorem 3, on the event \( \{ \lambda_{\max}(P_{n,r}^{-1}) \to \infty \text{ and } \xi_{n,r}^T P_n \xi_{n,r} \to 0 \text{ for } r = 1, \ldots, \nu \} \),

\[
(6.6) \quad \sum_{i=1}^{n} \| \tilde{y}_{i+d} - \tilde{y}_{i+d} \|^2 \leq (2d - 1) \sum_{r=1}^{\nu} \{ \lim_{i \to \infty} E(\eta_i^2(r) | F_{i-d}) + o(1)) \} \log \det (P_{n,r}^{-1})
\]

\[
+ o(\sum_{i=1}^{n} \rho_i) + o(\sum_{j:n(j) \leq n+d} \sum_{r=1}^{\nu} \sum_{s=1}^{h+d-1} [\| \xi_{n(j), r-s} \|^2 + \| \xi_{n(j), -s} \|^2]) \quad \text{a.s.}
\]

As pointed out in §1, the global convergence property (1.22) of adaptive \( \Delta \)-step ahead predictors associated with the stochastic gradient algorithm (1.24) in the case \( \Delta = 1 \), or (1.29) in the case \( \Delta > 1 \), has played a basic role in the development of self-optimizing rules by Goodwin et al. [14], [15] in the adaptive control problem of setting the input \( u_t \) at stage \( t \) so that the output \( y_{t+\Delta} \) is as close as possible to some target value \( y_{t+\Delta}^* \). For the case \( \Delta = 1 \), the property (4.1) of the adaptive 1-step ahead predictors \( \hat{y}_{n+1} = \Theta'_n \Phi_n \) based on the AML algorithm (1.13) has led to the construction by Lai and Wei [9] of adaptive control rules that satisfy

\[
(6.7) \quad \sum_{i=2}^{n} (y_i - y_i^* - \epsilon_i)^2 = O(\log n) \quad \text{a.s.,}
\]

which is a much stronger conclusion than the self-optimizing property (1.23), under the positive real condition (1.15) on \( C(z) \) and stability assumptions on \( A(z) \) and \( B(z) \).
In the unit-delay case \((\Delta = 1 = d)\), \(\eta_i = \epsilon_i\) and \(\tilde{y}_i = y_i - \epsilon_i\), so (6.7) follows from (4.1) if the input \(u_i\) is chosen by the certainty-equivalence equation (1.25) and if it can be shown that

\[
(6.8) \quad \limsup_{n \to \infty} \Phi_n' P_n \Phi_n < 1 \quad \text{and} \quad \log \det (P_n^{-1}) = O(\log n) \quad \text{a.s.}
\]

However, it is not even clear that (1.25) is well defined since the component \(b_1^{(t)}\) (estimating \(b_1\)) of \(\Theta_t\) may be 0 unless some continuity assumptions are made on the distribution of \(\{\epsilon_n\}\). To circumvent this difficulty and to ensure that (6.8) holds, Lai and Wei [9] suggested using blocks of white-noise probing inputs to improve the information content of the data whenever \(b_1^{(t)} = 0\) or \(\text{tr}(P_t)\) falls below some prespecified threshold \(\delta_t\) that decreases to 0 as \(t \uparrow \infty\). Using these occasional white-noise probing inputs in conjunction with the certainty-equivalence equation (1.25) associated with the AML algorithm, they were able to show that the resultant rule satisfies (6.7).

Despite the much stronger property (6.7) of these occasionally disturbed extended least squares certainty-equivalence rules than the self-optimizing property (1.23), the requirement that \(A(z)\) be stable in [9] imposes a serious limitation, particularly in light of the fact that the stochastic gradient certainty-equivalence rules introduced by Goodwin et al. [14], [15] can be used to stabilize a system that would be unstable in the open loop. Moreover, the results of [9] are only relevant to the unit-delay case. The results of §§3 and 5 have recently enabled us remove these limitations and to provide an asymptotically efficient adaptive control rule in the general delay case and without stability assumptions on \(A(z)\). The rule involves parallel implementation of the stochastic gradient and monitored recursive maximum likelihood algorithms. The stochastic gradient component of the controller serves to stabilize the system even when \(A(z)\) is not stable, as an application of Theorem 1. Together with an occasional dither signal to perturb the target values, it also leads to well excited blocks of input-output data from which strongly consistent estimates of the system parameters can be obtained by the method of moments to guide the recursive maximum likelihood algorithm, giving a control rule that can be shown by an application of Theorem 3 to satisfy

\[
(6.9) \quad \limsup_{n \to \infty} \sum_{i=\Delta+1}^{n} (y_i - \hat{y}_i - \eta_i)^2 / \log n \leq (2d-1)(p(d)+k+d-1+h) \limsup_{i \to \infty} E(\eta_i^2 | F_{i-d}) \quad \text{a.s.,}
\]

in the general delay case and without assuming \(A(z)\) to be stable. The details are given in [24].
REFERENCES


