A REPRESENTATION THEOREM FOR THE ERROR OF RECURSIVE ESTIMATORS

László Gerencsér
MTA SZTAKI
H-1111 Budapest, Kende 13-17, Hungary.

Abstract

The ultimate objective of this paper is to develop new techniques that can be used for the analysis of performance degradation due to statistical uncertainty for a wide class of linear stochastic systems. For this we need new technical tools similar to those used in [24] and summarized in Section 2. The immediate technical objective is to extend the technical results used in [24] to the Djereveckii-Fradkov-Ljung scheme with enforced boundedness, given as *Algorithm DFL*, (3.53)-(3.54), see [3, 13, 53]. Our starting point is a standard approximation of the estimation error used in the asymptotic theory of recursive estimation, see e.g. [54, 65]. Tight control of the difference between the estimation error and its standard approximation, referred to as *residuals*, is a crucial point in our applications. The main technical advance of the present paper is a set of strong approximation theorems for three closely related recursive estimation algorithms, given as Theorems 4.1-4.3, in which, for any $q \geq 1$, the L_q-norms of the residual terms are shown to tend to zero with rate $N^{-1/2-\varepsilon}$ with some $\varepsilon > 0$. This is a significant extension of previous results for the recursive prediction error or RPE estimator of ARMA processes given in [19] and [22]. Two useful corollaries will be derived in Section 5 and 6. In Theorem 5.1 a standard transform of the estimation-error process for the basic recursive estimation method, Algorithm CR, will be shown to be L-mixing, while in Theorem 6.2 the asymptotic covariance-matrix of the estimator for the same method will be given. Applications to multi-variable adaptive prediction and the minimum-variance self-tuning regulator for ARMAX-systems will be described in Section 7.
Keywords: adaptive prediction; stochastic complexity; recursive estimation; L-mixing processes; asymptotic covariance; stochastic adaptive control.

1 Introduction

The ultimate objective of this paper is to develop new techniques for the analysis of performance degradation due to statistical uncertainty for a wide class of linear stochastic systems. Performance degradation due to statistical uncertainty, called *regret*, following [46], can be computed at a single time moment, yielding instantaneous regret, or it can be summed over time, yielding cumulative regret. The objective of the paper is to develop new techniques that can be used for analyzing the pathwise (almost sure) asymptotics of the cumulative regret for a class of adaptive prediction and stochastic adaptive control problems.

A number of examples on the interaction of identification and control are available in the *identification for control* literature, see [29, 40, 41]. While the above papers contain fundamentally new ideas, the analysis that they present contains heuristic elements. In particular, they assume the independence of actually weakly dependent quantities in order to simplify the computation of the instantaneous regret. The present paper lays the foundations for a rigorous discussion of these heuristic arguments. Special examples of these new technical tools have been developed in the context of adaptive prediction of ARMA-processes in [24].

The immediate *technical objective* is a detailed analysis of the Djereveckii-Fradkov-Ljung scheme with enforced boundedness, given as *Algorithm DFL*, (3.53)-(3.54). This is a practically useful recursive estimation method introduced in [11, 12, 51] with a wide range of applications, see [3, 53]. The algorithm in its original form is given under (3.50) and (3.51) which is a potentially divergent procedure. To ensure convergence the original method is modified by enforced boundedness, a device that has been widely used in practice and rigorously analyzed in [19]. The study of the DFL scheme is reduced to the study of two related stochastic approximation methods Algorithm DR (discrete-time recursion) and Algorithm CR (continuous-time recursion),

described in Section 3. The conditions under which these methods are analyzed are very close to what we had in [19]. However, a critical condition imposed on the initialization of the process has been significantly simplified. Our conditions will be compared with other conditions used in the literature, in particular with those of [3], with emphasis on the so-called "boundedness-condition".

Asymptotic properties of recursive estimation processes are established in classical theory by using a series of approximations (see e.g. [54]). Thus we get a standard approximation of the error term, see e.g. [65] for a lucid exposition, for which limit results are easily established. Tight control of the difference between the estimation error and its standard approximation, that will be referred to as *residuals*, is crucial in the analysis of performance degradation due to statistical uncertainty, see [24].

The *main technical advance* of the present paper is a set of strong approximation theorems for three closely related recursive estimation algorithms, given as Theorems 4.1-4.3 of Section 4, in which, for any $q \geq 1$, the L_q-norms of the residual terms are shown to tend to zero with rate $N^{-1/2-\varepsilon}$ with some $\varepsilon > 0$. This is a significant extension of a previous result given in [19], where only the rate of convergence for the L_q-norms of the estimation error has been established and the explicit approximation of the estimation error and the residual term is not discussed at all. It extends also the result of [22] on the residual of the recursive prediction error estimator for ARMA-processes. The proof is quite demand in g: in addition to some basic inequalities developed in [17] the proof relies on [19] and uses a non-trivial moment inequality for weighted multiple integrals of L-mixing processes given in [21]. Preliminary versions of the results of Section 4 have been formulated in [20].

In comparison the material of Section 5 and 6 are relatively straightforward corollaries demand in g numerous small steps, though. In Theorem 5.1 a standard transform of the estimation-error process for the basic recursive estimation method, Algorithm CR, will be shown to be L-mixing, while in Theorem 6.2 the asymptotic covariance-matrix of the estimator for the same method will be given.

The *significance* of the results of the present paper is demonstrated by describing two applications in Section 7. In the first example the pathwise cumulative regret is quantified for an on-line adaptive predictor of multi-variable linear stochastic systems. In the second example a similar measure of performance degradation for the minimum-variance self-tuning regulator is computed. Both applications follow the arguments of [24], but heavily rely on the results of the present paper. A further application for indirect adaptive control of multi-variable linear stochastic systems is given in [27]. We think that the results are taylored to the needs of the users and they will pave the way to many further applications.

To motivate the studies carried out in this paper we will first give an illuminative application of less known technical results on off-line prediction error identification methods for ARMA processes. The application, given as Theorem 2.1, provides the answer to a basic problem of the theory of *stochastic complexity*, developed by Rissanen, see [58]: the performance degradation of adaptive predictors. The extension of this results to adaptive predictors using on-line estimation requires the extension of the relevant technical tools. First a strong approximation result for recursive prediction error identification methods for ARMA processes will be given as Theorem 2.4, this is also the starting point for the investigations of the present paper. Two important corollaries are Theorems 2.5 and 2.6. The relevance of these results in analyzing performance degradation in the context of on-line adaptive prediction of ARMA-processes will be described, culminating in Theorem 2.7. This theorem will be considered as a benchmark in future applications.

2 Adaptive prediction. Basic notions and conditions

An adaptive predictor for ARMA-processes is obtained if we use estimated systems-parameters in the prediction equation at time n as if it was the true value. Then we may ask, how much do we lose in prediction accuracy due to the inexact knowledge of the parameters. First we consider adaptive predictors using off-line estimation and indicate the nature of technical results

that are needed for the analysis. Then using the strong-approximation result (2.23) we arrive at analogous technical results for recursive estimation, which in turn can be applied to derive interesting properties of real-time adaptive predictors.

The set of real numbers will be denoted by $\mathbb{R}$, the p-dimensional Euclidean space will be denoted by $\mathbb{R}^p$. The Euclidean-norm of $x \in \mathbb{R}^p$ will be denoted by $|x|$. We shall often use subscripts to indicate partial derivatives.

Let (y_n), $0 \leq n < \infty$ be a wide-sense stationary ARMA (p, q) process satisfying the difference equation

$$\sum_{i=0}^{p} b_i^* y_{n-i} = \sum_{j=0}^{q} c_j^* e_{n-j},$$

or in shorthand notation $B^* y = C^* e$, where B^* and C^* are polynomials of the backward-shift operator of degree p and q, respectively. Define the polynomial $B^*(z^{-1}) = \sum_{i=0}^{p} b_i^*(z^{-i})$ and similarly $C^*(z^{-1})$. To estimate the system-parameters b_i^*, c_j^* from observed data (y_n) using the prediction error method the following technical assumption is assumed.

Condition 2.1 *The polynomials B^*, C^* are stable and relative prime, $b_0^* = c_0^* = 1$ and $b_p^* \neq 0, c_q^* \neq 0$.*

The condition $b_p^* \neq 0, c_q^* \neq 0$ has been assumed to allow the extension of our results to cases when the degree of one of the polynomials B^* or C^*, but not both is overestimated. The relevant work that we use is [1]. To characterize the noise-process we shall need the following definition that has been introduced in [17].

Definition 2.1 *We say that a discrete-time $\mathbb{R}^p$-valued stochastic process (u_n) is M-bounded if for all $1 \leq q < \infty$*

$$M_q(u) := \sup_{n \geq 0} \mathrm{E}^{1/q} |u_n|^q < \infty. \tag{2.1}$$

In this case we also write $u_n = O_M(1)$. For a stochastic process $(z_n), n \geq 0$ and a positive sequence (c_n) we write $z_n = O_M(c_n)$ if $u_n = z_n/c_n = O_M(1)$.

A basic tool that we will use is the theory of L-mixing processes, elaborated in [17] and used to solve some hard problems in system identification, see [18, 19, 22, 38, 43]. This concept is a generalization of what is called "exponentially stable processes" in the system-identification literature, see Definition 3.1 in 8.3 of [6] or [53]. We give the definition here for discrete time processes. Let a probability space $(\Omega, \mathcal{F}, P)$ be given together with a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+), n = 0, 1, ...$ such that (i) $\mathcal{F}_n \subset \mathcal{F}$ is monotone increasing (ii) $\mathcal{F}_n^+ \subset \mathcal{F}$ is monotone decreasing (iii) $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent for all n. For $n < 0$ we set $\mathcal{F}_n^+ = \mathcal{F}_0^+$.

Definition 2.2 *A stochastic process $u = (u_n), n = 0, 1, ...$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ if it is $\mathcal{F}_n$-adapted, M-bounded and with $\tau = 0, 1, ...$ and*

$$\gamma_q(\tau, u) = \gamma_q(\tau) = \sup_{n \geq \tau} \mathrm{E}^{1/q} |u_n - \mathrm{E}\ (u_n | \mathcal{F}_{n-\tau}^+)|^q,$$

we have

$$\Gamma_q = \Gamma_q(u) = \sum_{\tau=0}^{\infty} \gamma_q(\tau) < \infty. \tag{2.2}$$

The process u is L^+-mixing if in addition for all $q \geq 1$ there exist $C_q, c_q > 0$ such that for all non-negative integers τ

$$\gamma_q(\tau, u) \leq C_q (1 + \tau)^{-1 - c_q}.$$

Discussion of L-mixing. The prime example for L-mixing processes is a sequence of i.i.d. random variables with finite moments of all order. The response of an exponentially stable linear filter, with an L-mixing process as its input, is L-mixing. Products of L-mixing processes are also L-mixing. These properties make sure that the *verification* of L-mixing is typically easy in problems of system identification. The same invariance properties hold for the class of L^+-mixing processes. For "exponentially stable processes" we would require that $\gamma_q(\tau)$ converges to

0 geometrically fast, at least for some values of q, typically for $q = 4$. We shall need conditions for higher order moments to derive sharp bounds for the error terms in certain uniform laws of large numbers.

Condition 2.2 *The system-noise process* (e_n), $0 \le n < \infty$, *defined over an underlying probability-space* $(\Omega, \mathcal{F}, P)$, *is an* M*-bounded process. Moreover there is an increasing sequence of* σ*-fields* $(\mathcal{F}_n)$, $0 \le n < \infty$, $(\mathcal{F}_n) \subset \mathcal{F}$, *such that* (e_n) *is a martingale-difference process with constant conditional variance:*

$$\mathrm{E}(e_n|\mathcal{F}_{n-1}) = 0, \qquad \mathrm{E}(e_n^2|\mathcal{F}_{n-1}) = \sigma^2 = \text{const.}$$

almost surely. Finally, we assume that (e_n) *is* L*-mixing with respect to a pair of families of* σ*-algebras* $(\mathcal{F}_n, \mathcal{F}_n^+)$.

It follows that (e_n) is a wide-sense stationary orthogonal process. Conditions 2.1, 2.2 together will be called the *standard conditions* for ARMA-processes.

Discussion of the moment-condition. The difference between our conditions and conditions given in standard works such as [6] or [35] is that there only the condition

$$M_4(e) := \sup_{n \ge 0} \mathrm{E}^{1/4}|e_n|^4 < \infty.$$

is required. See the comment to Definition 3.1 in 8.3 of [6], or condition (4.1.20) of [35]. Thus our condition is much stronger, but our conclusions given in Theorem 2.2 will be also significantly stronger than the results of [6] and [35] given in terms of classical concepts such as strong consistency, central limit theorem, or the law of iterated logarithm, which all follow from our result and the corresponding result for martingales. M-boundedness could be relaxed by requiring the uniform boundedness of moments of *sufficiently high order*, however the order of the moments would depend on the order of the ARMA-process, i.e. on p and q. This is due to the fact that in our proof we rely on Kolmogorov's continuity theorem for random fields to get sharp bounds for the error terms in uniform laws of large numbers, which requires the existence of finite moments up to an order strictly greater than $(p + q)$ in the present application, see Theorem 8.3 of the Appendix.

Set $\theta^* = (b_1^*, ..., b_p^*, c_1^*, ..., c_q^*)^T$. Let $D_C \subset \mathbb{R}^q$ denote the set of vectors $(c_1, ..., c_q)$ such that the corresponding polynomial C^* is stable, let $D_B = \mathbb{R}^p$ and let

$$D_\theta = D_B \times D_C \subset \mathbb{R}^{p+q}.$$

Let $D_{\theta 0} \subset D_\theta$ be a compact domain such that $\theta^* \in \mathrm{int} D_{\theta 0}$. Then the *prediction-error method* for estimating the parameter θ^* is defined as follows (cf. e.g. [6, 35]): first take an arbitrary $\theta \in D_{\theta 0}$ and define an estimated prediction error process $\overline{\varepsilon} = (\overline{\varepsilon}_n(\theta))$ by the inverse equation

$$C\overline{\varepsilon} = By \tag{2.3}$$

using zero initial conditions. Define the cost-function

$$V_N(\theta) = \frac{1}{2} \sum_{n=1}^{N} \overline{\varepsilon}_n^2(\theta).$$

Minimizing $V_N(\theta)$ over $D_{\theta 0}$ yields an estimate $\widehat{\theta}_N$.

A precise definition of $\widehat{\theta}_N$ taking into account the possibility of the existence of several local minima can be given as follows: let $\Omega' \subset \Omega$ be a measurable set such that the equation

$$\frac{\partial}{\partial \theta} V_N(\theta) = 0$$

has a unique solution in the interior of $D_{\theta 0}$ denoted by $\mathrm{int} D_{\theta 0}$ on the event $\Omega' \subset \Omega$. Then this solution will be accepted as $\widehat{\theta}_N$ on Ω', while $\widehat{\theta}_N$ is defined as an arbitrary $D_{\theta 0}$-valued random

variable on $\Omega \setminus \Omega'$. It can be shown that we can take Ω' so that $P(\Omega') > 1 - C_q N^{-q}$ for any $q > 0$, see Lemma 2.1 [18], the proof of which is partially based on [1].

Remark. The uniqueness result of [1] remains valid if we redefine D_θ so that the degree of one of the polynomials B^* or C^*, but not both is overestimated. This is why $b_p^* \neq 0, c_q^* \neq 0$ has been assumed in Condition 2.1.

The quantity to be studied in the context of adaptive prediction is the prediction error $\overline{\varepsilon}_n(\widehat{\theta}_{n-1})$. We ask how much do we lose in prediction accuracy due to the statistical uncertainty present in $\widehat{\theta}_{n-1}$. A basic result says that, assuming that the standard conditions, Conditions 2.1, 2.2, are satisfied then the excess in mean prediction error, also called the *regret*, see [46], satisfies

$$\mathrm{E}(\overline{\varepsilon}_n^2(\widehat{\theta}_{n-1})) - e_n^2) = \frac{\sigma^2(e)}{n}(p+q)(1+o(1)). \tag{2.4}$$

This result is given in [24] and, under different conditions in [64]. It extends the result of [8] for AR-processes. A similar result for the cumulative regret for Gaussian linear regression was proved in [56] (see also Theorem 5.3 in [58]).

Summation over n in (2.4) gives that the left hand side is asymptotically equivalent to $\sigma^2(e)(p+q)\log N$. It has been shown in Theorem 1.1. of [24] that we can remove the expectation operator and we get the following *pathwise* result for the cumulative regret:

Theorem 2.1 *Assume that the standard conditions, Conditions 2.1, 2.2, are satisfied. Then*

$$\lim_{N\to\infty} \frac{1}{\log N} \lim_{N\to\infty} \sum_{n=1}^{N} (\overline{\varepsilon}_n^2(\widehat{\theta}_{n-1}) - e_n^2) = \sigma^2(e)(p+q) \quad \text{a.s.} \tag{2.5}$$

This result, under different conditions, was given for AR-processes in [36] and [37]. The much more difficult ARMA case was solved in [24] and [64], using different conditions and different methods. Note that classical limit theorems are not suitable to derive (2.5). Both results, (2.4) and (2.5) play prominent role in the theory of stochastic complexity. The quantity

$$C_{1,N} = \sum_{n=1}^{N} \overline{\varepsilon}_n^2(\widehat{\theta}_{n-1}) \tag{2.6}$$

is called a predictive stochastic complexity in [58].

Technical tools: strong approximations. Now we come to some technical details that are essential in the proof of the above results. A key point is a characterization of the estimation error process which is more accurate than previously known results. Define the asymptotic cost function by

$$W(\theta) = \lim_{n\to\infty} \frac{1}{2} \mathrm{E}\, \overline{\varepsilon}_n^2(\theta).$$

In the Gaussian case this is the asymptotic log-likelihood function modulo constants. It is easy to see that $W_\theta(\theta^*) = 0$, where $_\theta$ denotes differentiation with respect to θ. Also it is well-known that

$$R^* = W_{\theta\theta}(\theta^*) = \lim_{n\to\infty} \mathrm{E}\, \overline{\varepsilon}_{\theta n}(\theta^*)\overline{\varepsilon}_{\theta n}(\theta^*)^T$$

is nonsingular and in fact positive definite. Then we have the following representation of the estimation error (cf. [18]):

Theorem 2.2 *Assume that the standard conditions for ARMA-processes, Conditions 2.1, 2.2, are satisfied, then we have*

$$\widehat{\theta}_N - \theta^* = -(R^*)^{-1} \frac{1}{N} \sum_{n=1}^{N} \overline{\varepsilon}_{\theta n}(\theta^*) e_n + O_M(N^{-1}). \tag{2.7}$$

The *main contribution* of Theorem 2.2 is that the residual term has been shown to be of the order of magnitude $O_M(N^{-1})$. This is an improvement over the classical results of [50, 59] in the ARMA-case. The significance of this improvement is easily demonstrated: the residual term is sufficiently small so that limit theorems such as LIL and invariance principles for the estimator process can be immediately derived using martingale limit theory (cf. [33]). But the real motivation behind Theorem 2.2 had been the need to verify Rissanen's tail-condition for Gaussian ARMA-processes, introduced in the seminal paper [57], which in turn can be used to derive a lower bound for the cumulative loss in performance of any adaptive predictor for Gaussian ARMA-processes (cf. [26]).

Discussion of mixing conditions. There are a number of other notions of mixing. The best known notion is ϕ-mixing, an excellent and concise introduction to which is given in Chapter 7.2 of [15]. The measure of mixing for two σ-algebras $\mathcal{G}$ and $\mathcal{H}$ is defined for any $1 \leq p \leq \infty$ as follows:

$$\phi_p(\mathcal{G}|\mathcal{H}) = \sup_{A \in \mathcal{G}} ||P(A|\mathcal{H}) - P(A)||_p, \tag{2.8}$$

where $||\xi||_p$ denotes the L_p norm of the random variable ξ. It can be shown that for $p = 1$ we have $\frac{1}{2}\phi_1(\mathcal{G}|\mathcal{H}) = \alpha(\mathcal{G}, \mathcal{H})$, where

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{A \in \mathcal{G}, B \in \mathcal{H}} |P(AB) - P(A)P(B)| \tag{2.9}$$

is the familiar measure of strong mixing. Similarly, we have for $p = \infty$

$$\phi_\infty(\mathcal{G}|\mathcal{H}) = \sup_{A \in \mathcal{G}, B \in \mathcal{H}, P(B) > 0} |P(A|B) - P(A)| \tag{2.10}$$

which is the familiar measure of uniform mixing. A stochastic process (x_n) is then ϕ_p-mixing if with

$$\phi_p(n) = \phi_p(\mathcal{F}_n|\mathcal{F}_n^+),$$

where now $\mathcal{F}_n = \sigma\{x_i, i \leq n\},\ \mathcal{F}_n^+ = \sigma\{x_i, i \geq n\}$, we have $\lim_{n\to\infty} \phi_p(n)$.

In contrast to L-mixing, the *verification* of even the weakest form of ϕ-mixing, which is called for historical reasons strong mixing or α-mixing, is non-trivial even for Gaussian processes (see Chapter 17 of [42]). On the other hand, measurable static functions of ϕ-mixing processes are ϕ-mixing, while this may not be the case e.g. for discontinuous functions of L-mixing processes. (For a positive statement see [23] Theorem II.7).

¿From the point of view of *usefulness* both notions are equally useful for *off-line estimation*. Namely, the key technical device in analyzing off-line estimators is a kind of improved Hölder-inequality, see Lemma 8.1 of the Appendix, or Chapter 7.2 of [15], or Appendix III of [33]. In fact, it can be shown that the theorem remains valid even if the assumption that (e_n) is L-mixing is completely removed, since the remaining conditions imply the validity of certain improved Hölder-inequalities. The situation is quite different for recursive estimation methods, where L-mixing is heavily exploited. Further discussion on this will be given in Section 3.

The first step in the proof of Theorem 2.1 is to consider a second-order Taylor-series expansion of the terms on the left hand side. The estimation error process is handled using a standard transformation in the stochastic approximation literature. Define a piecewise constant continuous-time extension of $\widehat{\theta}_n$, and, denoting the time variable by t, introduce a new process by first normalizing $(\widehat{\theta}_t - \theta^*)$ to $t^{1/2}(\widehat{\theta}_t - \theta^*)$ and then using an exponential change of time scale $t = e^s$. Thus we get a new process

$$\psi_s = e^{s/2}(\widehat{\theta}_{e^s} - \theta^*). \tag{2.11}$$

A key observation is that the *transformed process* (ψ_s) is L-mixing with respect to $(\mathcal{F}_{e^s}, \mathcal{F}_{e^s}^+)$. For the definition of L-mixing in continuous time see the next section. The proof of this fact is based on Theorem 2.2 and the following simple result given as Theorem 3.3 in [24]:

Lemma 2.1 *Let $(u_t), t \geq 0$ be a zero-mean L-mixing process with respect to some pair of families of σ-algebras $(\mathcal{F}_t, \mathcal{F}_t^+)$. Let*

$$x_T = T^{-1/2} \int_1^T u_t dt. \tag{2.12}$$

Then the process $(y_s) = (x_{e^s})$ *is L-mixing with respect to the pair of families of* σ*-algebras* $(\mathcal{F}_{e^s}, \mathcal{F}^+_{e^s})$.

With this observation it can be shown that the process $\bar{\varepsilon}_n(\theta^*)$ and its gradient are asymptotically independent of $(\widehat{\theta}_N - \theta^*)$, which is exploited in proving (2.4). On the other hand it can be shown that the result given in (2.5) is essentially a law of large numbers in the new time-scale.

Now we come to the extensions of the above results for the case of on-line or *recursive estimation* of θ^*. The most widely used recursive estimation methods for ARMA-processes is the recursive prediction error (RPE) method, which in the case of Gaussian-processes reduces to the recursive maximum-likelihood (RML) method, see [6, 53]. This procedure serves as a *benchmark* for the general theory to be developed in Section 3, in particular it is a prime example for the Djereveckii-Fradkov-Ljung scheme or DFL scheme. Both for theoretical and practical reasons we consider recursive prediction error processes $\widehat{\widehat{\theta}}_n$ using a resetting mechanism to enforce the boundedness of the estimator. The convergence analysis for such a procedure has been given in Theorem 4.2 of [19].

We will now give the details of the RPE method for ARMA processes and a set of technical conditions that we use to guarantee convergence. The conditions are simpler than those given in Section 4 of [19]. We will shortly indicate how the present conditions imply the conditions given for the DFL scheme in the next section. Most of the discussion of these conditions will be deferred to the next section. The definition of the RPE-method *without resetting* is

$$\widehat{\widehat{\theta}}_n = \widehat{\widehat{\theta}}_{n-1} - \frac{1}{n}(\widehat{\widehat{R}}_{n-1})^{-1} \frac{\partial}{\partial\theta}\varepsilon_n \cdot \varepsilon_n \tag{2.13}$$

$$\widehat{\widehat{R}}_n = \widehat{\widehat{R}}_{n-1} + \frac{1}{n}((\frac{\partial}{\partial\theta}\varepsilon_n)(\frac{\partial}{\partial\theta}\varepsilon_n)^T - \widehat{\widehat{R}}_{n-1}) \tag{2.14}$$

with some initial conditions $(\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_1)$, where ε_n and $\frac{\partial}{\partial\theta}\varepsilon_n$ denote on-line estimates of $\bar{\varepsilon}_n(\theta^*)$ and of $\frac{\partial}{\partial\theta}\bar{\varepsilon}_n(\theta)|_{\theta=\theta^*}$. These are obtained by using the most recent estimations of B^* and C^* in the linear filters defining the current values of $\bar{\varepsilon}_n(\theta^*)$ and of $\frac{\partial}{\partial\theta}\bar{\varepsilon}_n(\theta)|_{\theta=\theta^*}$. Thus e.g. ε_n is defined by the time-varying filter

$$(\widehat{\widehat{C}}_{n-1}\varepsilon)_n = (\widehat{\widehat{B}}_{n-1}y)_n. \tag{2.15}$$

For further details see [53]. Note that, in contrast with [19], the recursion is initiated at time $n = 1$ rather than at time $n = 0$ to ensure a more convenient connection with continuous-time methods.

The RPE method without resetting is a special case of a general recursive estimation scheme, called the DFL scheme, to be described in details in the next section, see (3.50)-(3.51). Note that together with θ^* we also estimate the matrix R^*.

It is well-known from simulation examples that the RPE error method may diverge, unless some precaution is taken. This difficulty is often dealt with a controversial "boundedness condition" first formulated in [51]. This will be discussed in detail in the context of the DFL-method. A convergent *truncated* RPE method has been given in Section 4 of [19], which we now describe. Let

$$D_R = \mathbb{R}^+(p \times p), \quad \text{and} \quad D = D_\theta \times D_R,$$

where $\mathbb{R}^+(p \times p)$ denotes the set of symmetric, positive definite $p \times p$ matrices. Let $D_{\theta 0} \subset D_\theta$ be a compact set containing θ^* in its interior and similarly let $D_{R0} \subset D_R$ be a compact set containing R^* in its interior and let $D_0 = D_{\theta 0} \times D_{R0}$.

Resetting: If at any time n the next estimator $(\widehat{\widehat{\theta}}_{n+1}, \widehat{\widehat{R}}_{n+1})$ would leave D_0 then we redefine its value by resetting it to the initial value, i.e.

$$\text{for} \quad (\widehat{\widehat{\theta}}_{n+1}, \widehat{\widehat{R}}_{n+1}) \notin \text{int} D_0 \quad \text{reset as} \quad (\widehat{\widehat{\theta}}_{n+1}, \widehat{\widehat{R}}_{n+1}) := (\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_0). \tag{2.16}$$

To avoid being trapped to the boundary of the truncation domain the initial value $(\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_0)$ must be aligned to $D_{\theta 0} \times D_{R0}$, as described in Condition 3.4. This condition is given in terms of the

so-called associated ODE. Define for $\theta \in D_\theta$

$$R^*(\theta) = \lim_{n\to\infty} \mathrm{E}\bar{\varepsilon}_{\theta n}(\theta)\bar{\varepsilon}_{\theta n}(\theta)^T. \tag{2.17}$$

Then obviously $R^* = R^*(\theta^*)$. With this notation the associated ODE, with the time-variable v, is defined as

$$\begin{aligned} \dot{\theta}_v &= -R_v^{-1}\frac{\partial}{\partial\theta}W(\theta_v), \\ \dot{R}_v &= R^*(\theta_v) - R_v. \end{aligned} \tag{2.18}$$

The right hand side is defined in $D_\theta \times D_R$. This is the usual way of defining the associated ODE, see [3, 53]. However in [19] as well as later in this paper we will define the associated ODE by using a change of time scale $t = e^v$.

The condition ensuring that resetting works for the general recursive estimation methods given in Section 3, including the DFL scheme is Condition 3.4. Following the arguments of Section 4 of [19] it is easy to see that the first part of Condition 3.4, requiring a certain kind of asymptotic stability of the associated ODE, follows for the RPE method. Namely, it follows directly from [1] that (2.18) has a unique stationary point in $D_\theta \times D_R$, which is (θ^*, R^*). It is also easy to see that this equilibrium point is asymptotically stable, since the eigenvalues of the Jacobian-matrix of the right hand side of the ODE at (θ^*, R^*) are all -1. Now it is easy to show that the associated ODE is globally asymptotically stable in $D_\theta \times D_R$. For the proof we need the observation that $W(\theta_v)$ is non-increasing as long as R_v is positive definite and R_v is bounded and positive definite as long as θ_v belongs to a fixed compact set.

Let $x_n = (\widehat{\widehat{\theta}}_n, \widehat{\widehat{R}}_n)$ denote the estimator at time n, let $z = (\theta, R)$ denote a running parameter and let $z(v, u, \xi)$ denote the solution of (2.18) with initial value ξ at time u. Then it follows that for every $\xi \in D_0, v \geq u \geq 0$ the solution $z(v, u, \xi) \in D$ is defined for $1 \leq s \leq t < \infty$, it converges to $x^*(\theta^*, R^*)$ for $t \to \infty$ and we have with some C_0 and $\alpha = 1 - c$ with arbitrary small $c > 0$

$$\|\frac{\partial}{\partial\xi}z(v, u, \xi)\| \leq C_0 e^{-\alpha(v-u)}. \tag{2.19}$$

It follows, using a change of time-scale $t = e^v$, that the first part of Condition 3.4 is satisfied. Here $\|\cdot\|$ denotes the operator norm of a matrix.

To ensure the validity of the second part of Condition 3.4 we have to assume that some a priori knowledge of the system parameters θ^* and the Hessian R^*, say $\xi = (\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_0)$ are available. They can be obtained e.g. from an off-line estimation.

Condition 2.3 *Let $D_0 = D_{\theta 0} \times D_{R0} \subset D_\theta \times D_R$ be a compact truncation domain such that $x^* = (\theta^*, R^*) \in$ int D_0. (i) It is assumed that there exists a compact convex domain $D_0' \subset D$ such that*

$$z(v, u, x) \in D_0' \quad \text{for } x \in D_0 \quad \text{and } z(v, u, x) \in D \quad \text{for } x \in D_0' \quad \text{for all } v \geq u \geq 0. \tag{2.20}$$

(ii) It is assumed that we have an initial estimate $\xi = (\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_0)$ such that for any $v \geq u \geq 0$ we have $z(v, u, \xi) \in$ int D_0.

Remark. Since our objective is to restate Theorem 4.2 of [19], part (iii) of Condition 3.4 of the present paper need not be verified at this time, since it was not required in [19]; it is special addition for the present paper.

To ensure the stability of the time-varying filter (2.15) given as $(\widehat{\widehat{C}}_{n-1}\varepsilon)_n = (\widehat{\widehat{B}}_{n-1}y)_n$ we need a second condition imposed on the truncation domain (cf. Condition 3.7 of Section 3 given for the DFL method). Let us consider a fixed state-space realization of the inverse system (2.3) and let the state-transition-matrix be denoted by $\widetilde{C}$. In [19] this is given as the so-called companion-matrix corresponding to the polynomial C (see Condition 4.5 of [19]). Let $D_{B0} \subset D_B$ and $D_{C0} \subset D_C$ be compact domains and let

$$D_{\theta 0} = D_{B0} \times D_{C0}.$$

Now Condition 3.7 would read as follows:

Condition 2.4 *Let $D_{\tilde{C}0}$ denote the set of matrices $\tilde{C}$, when C is taken from D_{C0}. Then $D_{\widetilde{C}0}$ is jointly stable in the sense that there exists a single $q \times q$ symmetric positive definite matrix U and $0 < \lambda < 1$ such that for all $\tilde{C} \in D_{\tilde{C}0}$*

$$\widetilde{C}^T U \widetilde{C} \ \leq \ \lambda U.$$

It follows that there exists some $c > 0$ such that for any sequence $(\tilde{C}_n)$ with $\tilde{C}_n \in D_{\tilde{C}0}$ we have

$$||\tilde{C}_n ... \tilde{C}_0|| \leq c\lambda^{n/2}. \tag{2.21}$$

A discussion of the joint stability condition. Condition 2.4 above is required only to ensure that (2.21) holds. In the system-identification literature it had been occasionally implicitly assumed that the individual stability of each $\tilde{C} \in D_{\tilde{C}0}$ implies (2.21), see e.g. [34]. This is easily seen to be wrong. One way to ensure Condition 2.4 is to choose the truncation domain D_0 small, but this is obviously not practical. A better way is to use a suitable realization of the inverse system (2.3). To indicate the potential of alternative realizations let us consider a Gilbert-Kalman realization of the inverse system (see [44]). Assume that the roots of the polynomial $C = C(z^{-1})$ are all real and simple and let them be denoted by λ_i. Then we will have

$$\tilde{C} = \mathrm{diag}(\lambda_i),$$

and obviously any compact set of matrices $D_{\tilde{C}0}$ is jointly stable. Potentially useful alternative realizations are given in [55]. A second way of ensuring the validity of (2.21) is given in [3]. This will be discussed in connection with the DFL scheme in the next section.

Finally we will need two additional conditions for the noise process. First, the M-boundedness of (e_n) is further strengthened by assuming the existence and boundedness of certain exponential moments.

Condition 2.5 *We assume that $|e_n|^2$ is in class M^*, i.e for some $\varepsilon > 0$ we have*

$$\sup_n \mathrm{E} \ \exp \varepsilon |e_n|^2 < \infty.$$

This condition is certainly satisfied if (e_n) is a stationary Gaussian process. The role of this condition will be discussed in Section 3 in the context of the DFL scheme.

Secondly, we need to be more specific on the mixing rate of (e_n). The condition to follow is motivated by Lemma 3.1 [24], which states for continuous-time L-mixing processes (cf. Definition 3.2) that, if (u_t) is an L-mixing process then, $\gamma_q(\tau, u) \leq 4\Gamma_q(u)/\tau$ for all $q \geq 1$ and $\tau \geq 0$. The validity of a slightly stronger inequality is required by the following condition in discrete time (cf. Condition 3.9 of Section 3 given for the DFL scheme):

Condition 2.6 *We assume that (e_n) is L^+-mixing with respect to a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+)$.*

The role of this condition is in the analysis of the difference between the "frozen parameter" process $\bar{\varepsilon}_n(\theta)$ evaluated at $\theta = \widehat{\theta}_n$ and its on-line estimate ε_n, see [19], Lemma 5.6 restated as Lemma 3.2 of the present paper. From the purely technical point of view, L^+-mixing is used in [19], Theorem 6.1. In view of the general theorem for the DFL scheme, given as Theorem 3.3, we get the following result (see also Theorem 4.2 of [19]) :

Theorem 2.3 *Let (y_n) be an ARMA-process satisfying the standard conditions, Conditions 2.1, 2.2. Consider the recursive prediction error estimator defined by (2.13), (2.14), modified by a resetting mechanism given under (2.16). Let the truncation domain be of the form*

$$D_0 = D_{\theta 0} \times D_{R0} \quad \text{with} \quad D_{\theta 0} = D_{B0} \times D_{C0}.$$

Assume that D_0 satisfies Condition 2.3 and D_{C0} satisfies Condition 2.4. Finally let the innovation process satisfy the additional conditions Condition 2.5 and 2.6. Then for the recursive estimators $(\widehat{\widehat{\theta}}_N, \widehat{\widehat{R}}_N)$ we have

$$\widehat{\widehat{\theta}}_N - \theta^* = O_M(N^{-1/2}) \quad \textit{and} \quad \widehat{\widehat{R}}_N - R^* = O_M(N^{-1/2}). \tag{2.22}$$

One of the special features of this result is that the *moments* of the estimation error are bounded from above. While the above theorem is certainly of interest, it is obviously much weaker than the characterization of the off-line estimator given in Theorem 2.2. But Theorem 2.3 is a key technical tool in deriving a strong approximation theorem relating the recursive prediction error estimator to the off-line prediction error estimator. This result is given in [22], stating that under the conditions of Theorem 4.2 of [19] (and thus under the conditions of Theorem 2.3) we have

$$\widehat{\widehat{\theta}}_N - \widehat{\theta}_N = O_M(\frac{\log N}{N}). \tag{2.23}$$

Combining (2.23) with Theorem 2.2 we get:

Theorem 2.4 *Under the conditions of Theorem 2.3 we have*

$$\widehat{\widehat{\theta}}_N - \theta^* = -(R^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\overline{\varepsilon}_{\theta n}(\theta^*)e_n + O_M(\frac{\log N}{N}). \tag{2.24}$$

This strong approximation result provides a very precise characterization of $\widehat{\widehat{\theta}}_N$. The control of moments of the residual term is an essential feature of the result that is very much exploited in deriving Theorems 2.7 and 2.7. A direct corollary of the above theorem is the following:

Theorem 2.5 *Under the conditions of Theorem 2.3 we have*

$$\mathrm{E}N(\widehat{\widehat{\theta}}_N - \theta^*)(\widehat{\widehat{\theta}}_N - \theta^*)^T = \sigma^2(R^*)^{-1} + O(N^{-1/2}\log N). \tag{2.25}$$

Finally, taking into account Theorem 2.2, (2.23) and Lemma 2.1 we get:

Theorem 2.6 *Under the conditions of Theorem 2.3 the transformed process*

$$\psi_s = e^{s/2}(\widehat{\widehat{\theta}}_{e^s} - \theta^*) \tag{2.26}$$

is L-mixing with respect to $(\mathcal{F}_{e^s}, \mathcal{F}^+_{e^s})$.

The above three results, Theorems 2.4, 2.5 and 2.6, are the key tools in extending Theorem 2.1 to adaptive predictors using recursive estimators rather than off-line estimators (cf. [24]). Thus we get the following *key result*:

Theorem 2.7 *Under the conditions of Theorem 2.3 we have*

$$\lim_{N\to\infty}\frac{1}{\log N}\lim_{N\to\infty}\sum_{n=1}^{N}(\overline{\varepsilon}_n^2(\widehat{\widehat{\theta}}_{n-1}) - e_n^2) = \sigma^2(e)(p+q) \quad \text{a.s.} \tag{2.27}$$

In addition, the above proposition remains valid, if we replace $\overline{\varepsilon}_n(\widehat{\widehat{\theta}}_{n-1})$ by its on-line computed approximation ε_n, see (2.15):

Theorem 2.8 *Under the conditions of Theorem 2.3 we have*

$$\lim_{N\to\infty}\frac{1}{\log N}\lim_{N\to\infty}\sum_{n=1}^{N}(\varepsilon_n^2 - e_n^2) = \sigma^2(e)(p+q) \quad \text{a.s.} \tag{2.28}$$

The main *contribution* of the present paper is the extension of the technical result given as Theorems 2.4, 2.5 and 2.6 to general recursive estimation schemes that include the DFL scheme with enforced boundedness, given as (3.53)-(3.54). The extension of Theorem 2.4 uses the results of [17] and [19] but requires an additional technical tool given in [21]. This extension will be given in Section 3. The extensions of Theorems 2.5 and 2.6 are obtained using straightforward, though numerous approximations in Section 5 and 6. The present paper actively uses the results of [17, 19] and [21]. To facilitate reading, these relevant results are summarized in the Appendix.

3 General recursive estimation schemes

The prime objective of this section is to formulate a general recursive estimation method, the Djereveckii-Fradkov-Ljung scheme or DFL scheme with enforced boundedness, together with conditions that ensure its convergence. It is given as Algorithm DFL under (3.53)-(3.54), developed in [11, 12, 51], see also the books [3, 13, 53].

But first we present two closely related recursive algorithms: Algorithm CR (continuous-time recursion), (3.16) and Algorithm DR (discrete-time recursion), (3.34), which can be interpreted as "frozen parameter" approximations to the DFL scheme. The main results of the paper will be formulated and proved for the continuous-time method, Algorithm CR. The connection between the continuous-time and the discrete-time algorithm is straightforward. In contrast, the connection between Algorithm DR and the DFL scheme is not straightforward at all, but it has been worked out in [19], Section 6 and 7. Details will be given while discussing the DFL method.

Our first tentative general method is a continuous-time recursive estimation process without resetting, given by a random differential equation of the form

$$\dot{x}_t = \frac{1}{t}(H(t, x_t, \omega) + \delta H(t, \omega)), \qquad x_1 = \xi, \tag{3.1}$$

defined over the underlying probability-space $(\Omega, \mathcal{F}, P)$. Here x_t indicates an estimator sequence and $H = (H(t, x, \omega))$ is a random field defined in $[1, \infty) \times D$, where D is a bounded open domain in $\mathbb{R}^p \times \Omega$ and $\delta H(t, \omega))$ is a perturbation term to be described later. The advantage of continuous time is that some calculations can be carried out more easily than in discrete time.

The technical conditions that we impose on $H(t, x, \omega)$ will be tuned to fit the DFL scheme, given by (3.53) and(3.51) below. A continuous-time example for a random field $H(t, x, \omega)$ that is motivated by the DFL scheme is the following:

$$H(t, x, \omega) = \varepsilon(t, x, \omega)\eta(t, x, \omega), \tag{3.2}$$

where $\varepsilon(t, x, \omega)$ and $\eta(t, x, \omega)$ are stationary, jointly Gaussian-processes, defined by finite-dimensional stable linear filters applied to a standard Wiener-process (w_s):

$$\varepsilon(t, x, \omega) = \int_{-\infty}^{t} h_\varepsilon(t - s, x) dw_s, \qquad \eta(t, x, \omega) = \int_{-\infty}^{t} h_\eta(t - s, x) dw_s, \tag{3.3}$$

such that in an appropriate state-space representation the state-space matrices corresponding to the impulse-responses $h_\varepsilon(\tau, x)$ and $h_\eta(\tau, x)$ are sufficiently smooth functions of the parameter x. In the recursive maximum-likelihood identification method for discrete-time Gaussian-ARMA-processes $\varepsilon(n, x, \omega)$ would be the estimated input noise, with x being the system-parameter and $\eta(n, x, \omega)$ would be its negative gradient with respect to x, assuming stationary initialization for both processes. To specify the conditions to be imposed we need some preliminary technical details. The notion of M-bounded processes will now be extended to parameter-dependent, continuous-time processes.

Definition 3.1 *Let $D_0 \subset \mathbb{R}^p$ be a compact set and let $(u_t(x))$ be an $\mathbb{R}^k$-valued measurable stochastic process defined on $\Omega \times \mathbb{R}^+ \times D_0$, where $\mathbb{R}^+ = \{t : t \geq 0\}$. We say that $(u_t(x))$ is M-bounded (in D_0) if for all q with $1 \leq q < \infty$ we have*

$$M_q(u) = \sup_{\substack{t \geq 0 \\ x \in D_0}} \mathrm{E}^{1/q} |u_t(x)|^q < \infty. \tag{3.4}$$

If $(u_t(x))$ is M-bounded then we write $= O_M(1)$. We shall use the same terminology if x or t degenerate into a single point. If c_t is a sequence of positive numbers then we write $u_t(x) = O_M(c_t)$ if $u_t(x)/c_t = O_M(1)$.

The notion of L-mixing will now be extended to parameter-dependent, continuous-time processes. Let a probability space $(\Omega, \mathcal{F}, P)$ be given together with a pair of families of σ-algebras $(\mathcal{F}_t, \mathcal{F}_t^+)$ such that (i) $\mathcal{F}_t \subset \mathcal{F}$ is monotone increasing (ii) $\mathcal{F}_t^+ \subset \mathcal{F}$ is monotone decreasing and $\mathcal{F}_t^+$ is right continuous in t i.e. $\mathcal{F}_s^+ = \sigma\{\bigcup_{0 < \varepsilon} \mathcal{F}_{s+\varepsilon}^+\}$ (iii) $\mathcal{F}_t$ and $\mathcal{F}_t^+$ are independent for all t. For $s < 0$ we set $\mathcal{F}_s^+ = \mathcal{F}_0^+$.

Definition 3.2 *Let $D_0 \subset \mathbb{R}^p$ be a compact set and let $(u_t(x))$ be an $\mathbb{R}^k$-valued measurable stochastic process defined on $\Omega \times \mathbb{R}^+ \times D_0$. We say that $u = (u_t(x))$ is L-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in x for $x \in D_0$, if it is $\mathcal{F}_t$-progressively measurable, M-bounded (in D_0) and if for all $q \geq 1$ with*

$$\gamma_q(\tau, u) = \gamma_q(\tau) = \sup_{\substack{t \geq \tau \\ x \in D_0}} \mathrm{E}^{1/q} |u_t(x) - \mathrm{E}\ (u_t(x) | \mathcal{F}_{t-\tau}^+)|^q, \qquad \tau \geq 0,$$

we have

$$\Gamma_q = \Gamma_q(u) = \int_0^\infty \gamma_q(\tau) d\tau < \infty. \tag{3.5}$$

We say that $(u_t(x)), t \geq 0 \quad x \in D_0$ is L^+-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in x for $x \in D_0$, if in addition for all $q \geq 1$ there exist $C_q, c_q > 0$ such that for all $\tau \geq 0$

$$\gamma_q(\tau, u) \leq C_q (1 + \tau)^{-1 - c_q}. \tag{3.6}$$

The definition extends to parameter-free processes (u_t) and to discrete-time processes $(u_n(x))$. In the latter we set

$$\Gamma_q = \Gamma_q(u) = \sum_{\tau=0}^\infty \gamma_q(\tau) d\tau < \infty. \tag{3.7}$$

Condition 3.1 *The process $H = (H(t, x, \omega))$ is assumed to be defined in $\Omega \times \mathbb{R}^+ \times D$, where $D \subset \mathbb{R}^p$ is an open domain, it is three times continuously differentiable with respect to x for $x \in D$ almost surely and for any compact set $D_0 \subset D$ H and its derivatives up to order 3 are M-bounded in D_0. Furthermore $(H(t, x, \omega))$ and its first derivative $H_x = (H_x(t, x, \omega))$ are L^+-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in $x \in D_0$.*

In [19] we used the finite difference field of $H = (H(t, x, \omega))$, to capture the smoothness of $H(t, x, \omega)$. In general, we considered the process

$$\Delta u / \Delta x\ (t, x, x + h, \omega) = |u_t(x + h) - u_t(x)| / |h|$$

defined for $t \geq 0, x \neq x + h \in D$. We say that $u = (u_t(x))$ is *M-Lipschitz-continuous* with respect to x in D_0, if the process $\Delta u / \Delta x$ defined above is M-bounded, i.e. if for all $1 \leq q < \infty$ we have

$$M_q(\Delta u / \Delta x) = \sup_{\substack{t \geq 0 \\ x \neq x + h \in D_0}} \mathrm{E}^{1/q} |u_t(x + h) - u_t(x)|^q / |h| < \infty.$$

Condition 1.1. of [19] is then following:

Condition H. The process $(H(t, x, \omega))$ and $(\Delta H / \Delta x(t, x, x + h, \omega))$ are assumed to be separable and L^+-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in $x, x + h \in D$.

It is easy to see that Condition H is implied by Condition 3.1.

A discussion of L-mixing. We give further details for comparing L-mixing and ϕ-mixing as described in Chapter 7.2 of [15]. In L-mixing we consider projections on the relative future defined by $\mathcal{F}_{t-\tau}^+$ and the resulting approximation error is

$$\mathrm{E}^{1/q} |u_t - \mathrm{E}\ (u_t | \mathcal{F}_{t-\tau}^+)|^q \leq \gamma_q(\tau, u) \tag{3.8}$$

for $\tau \geq 0$. In ϕ-mixing we consider projections on the past and the corresponding error from the mean, defined as

$$||P(A | \mathcal{H}) - P(A)||_p \tag{3.9}$$

for $A \in \mathcal{G}$. Assuming that there is a random variable Φ such that

$$||P(A | \mathcal{H}) - P(A)|| \leq \Phi \tag{3.10}$$

for all $A \in \mathcal{G}$ we have the following proposition (see Proposition 2.6, (2.23) of Chapter 7.2 of [15]): let Z be a $\mathcal{G}$-measurable random variable such that $||Z||_s$ is finite and let $r, s > 1$ be such that $r^{-1} + s^{-1} = 1$. Then

$$||\mathrm{E}(Z|\mathcal{H}) - \mathrm{E}(Z)||_p \le 2||\Phi||_p^{1/r}||\mathrm{E}(|Z|^s|\mathcal{H})||_p^{1/s}. \tag{3.11}$$

Now if $(u_t), t \ge 0$ is L-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, then taking the conditional expectation of $u_t - \mathrm{E}\ (u_t|\mathcal{F}_{t-\tau}^+)$ with respect to $\mathcal{F}_{t-\tau}$ (cf. (3.8)), we get by Jensen's inequality and the assumed independence of $\mathcal{F}_{t-\tau}$ and $\mathcal{F}_{t-\tau}^+$

$$\mathrm{E}^{1/q}|\mathrm{E}\ (u_t|\mathcal{F}_{t-\tau}) - \mathrm{E}\ (u_t)|^q \le \gamma_q(\tau, u). \tag{3.12}$$

In this respect the two notions of mixing lead to similar conclusions.

We will need to strengthen the condition on the M-boundedness of H_x as follows for reasons that will be discussed later, following Condition 3.8.

Condition 3.2 *$H(t, x, \omega)$ is piecewise continuous in t almost surely and for any compact set $D_0 \subset D$ there exists a random variable $L_t = L_t(\omega) \ge 0$ such that for all $x \in D_0$*

$$|H_x(t, x, \omega)| \le L_t(\omega)$$

and here L_t is in class M^*, i.e. for some $\varepsilon > 0$ we have

$$\sup_t \mathrm{E} \exp(\varepsilon L_t) < \infty. \tag{3.13}$$

In [19] we had a weaker condition (see Condition 1.2 of [19]):

Condition L $H(t, x, \omega)$ is piecewise continuous in t and for any compact set $D_0 \subset D$ Lipschitz-continuous in x for $x \in D_0$ almost surely with a (t, ω)-dependent Lipschitz constant $L_t = L_t(\omega) \ge 0$, i.e. for $x, x' \in D_0$ we have

$$|H(t, x, \omega) - H(t, x', \omega)| \le L_t(\omega)|x - x'|,$$

where L_t is in class M^*.

Assuming that $(\delta H(t, \omega))$ is piecewise continuous in t almost surely, a solution (x_t) of (3.1) exists almost surely in some finite or infinite interval. A central role in the analysis of (x_t) is played by the mean-field of $H(t, x, \omega)$. To simplify the presentation it is assumed that the mean-field is essentially independent of t, but a small perturbation is allowed: we have $\mathrm{E}H(t, x, \omega) = G(x) + \delta G(t, x)$, where $\delta G(t, x)$ is small in a sense to be specified below.

Condition 3.3 *We have for any compact set $D_0 \subset D$ and $t \ge 0,\ x \in D_0$*

$$\mathrm{E}H(t, x, \omega) = G(x) + \delta G(t, x),$$

where $\delta G(t, x) = O(t^{-1/2-\varepsilon})$ uniformly in $x \in D_0$, with some $\varepsilon > 0$. $G(y)$ has continuous and bounded partial derivatives up to third order. Finally, we assume that

$$G(x) = 0 \tag{3.14}$$

has a unique solution x^* in D.

Remark. In [19] the slightly weaker condition $\delta G(t, x) = O(t^{-1/2})$ has been used (see Condition 1.3 of [19]). Also only differentiability up to order 2 was required.

Let us now consider the ordinary differential equation, the so-called associated ODE:

$$\dot{y}_t = \frac{1}{t} G(y_t), \qquad y_s = \xi,\ s \ge 1. \tag{3.15}$$

Under the condition above (3.15) has a unique solution in some finite or infinite interval, which we denote by $y(t, s, \xi)$. It is well-known that $y(t, s, \xi)$ is a twice continuously differentiable

function of ξ. The celebrated ODE-principle states that the solution trajectories of the random differential equation (3.1), under additional conditions, follow the solution trajectories of the associated ODE (3.15).

Interpreting (3.1) as a continuous-time stochastic approximation method for solving the non-linear algebraic equation $G(x) = 0$ an obvious difference compared to classical theory, (see [54]), is that G is not defined on the whole space. Thus we are lead to the study of recursive estimation methods *constrained* to a fixed domain D. In fact for theoretical reasons it is better to assume that the estimator process is constrained to a *compact* domain $D_0 \subset D$. One way to enforce boundedness of the estimation process is to restart it whenever it would leave D_0. Such a *truncated* version of (3.1) is described by Algorithms CR below, following [19]. A short discussion on the resetting mechanism to follow will be given in the context of the DFL scheme.

Algorithm CR. Consider a continuous-time recursion given by a random differential equation

$$\dot{x}_t = \frac{1}{t}(H(t, x_t, \omega) + \delta H(t, \omega)), \qquad x_1 = \xi \tag{3.16}$$

combined with the following *resetting* mechanism. Let $D_0 \subset D$ denote a compact truncation domain such that $x^* \in \text{ int } D_0$. Let us initialize (3.16) at some time $\sigma \geq 1$ and let $x_\sigma = \xi \in \text{ int} D_0$. Let

$$\tau(\sigma) = \min\{t : t > \sigma, x_t \in \partial D_0\}, \tag{3.17}$$

where ∂D_0, denotes the boundary of D_0. Then we reset x to $x_1 = \xi$, which is formally stated by requiring that the right hand side limit of x_t at $t = \tau = \tau(\sigma)$ will be ξ:

$$x_{\tau+} = \xi. \tag{3.18}$$

Thus we get a piecewise continuous trajectory (x_t) defined in some finite or infinite interval.

Remark. An alternative resetting mechanism, used in the analysis of discrete time processes, is obtained by putting

$$x_t = \xi \quad \text{for} \quad n < t \leq n+1 \ \text{ if } \ x_\tau \in \partial D_0 \ \text{ for } \ n < \tau \leq n+1 \tag{3.19}$$

To ensure that the estimator sequence is not bounced back and forth by resetting we need to impose some condition on the shape and relative position of the truncation domain, x^* and ξ, which is captured via the flow induced by the ODE. For this we need to define the star-like closure of the set D_0, relative to x^*, as follows:

$$D_0^* = \{y : y = x^* + \lambda(x - x^*),\ 0 \leq \lambda \leq 1, x \in D_0\}.$$

The condition below is a simplified and corrected version of Condition 1.5. of [19]. The simplification is that the condition on the position of the initial value $x_1 = \xi$ has been relaxed, while the correction is that an additional compact convex set D_0' containing the truncation domain has been introduced that has been implicitly used in the final step of the proof of Theorem 1.1. of [19], see (2.10) of [19].

Condition 3.4 *Let $D_0 \subset D$ be a compact truncation domain such that $x^* \in \text{int} D_0$. We assume (i) there exists a compact convex set $D_0' \subset D$ such that*

$$y(t, s, \xi) \in D_0' \quad \text{for } \xi \in D_0 \ \text{ and } y(t, s, \xi) \in D \ \text{ for } \xi \in D_0' \ \text{ for all } t \geq s \geq 1. \tag{3.20}$$

In addition $\lim_{t\to\infty} y(t, s, \xi) = x^*$ for $\xi \in D$ and

$$\|\frac{\partial}{\partial \xi} y(t, s, \xi)\| \leq C_0 (s/t)^\alpha. \tag{3.21}$$

with some $C_0 \geq 1, \alpha > 0$ for all $\xi \in D_0'$ and $t \geq s \geq 1$. (ii). We have an initial estimate $\xi = x_1$ such that for all $t \geq s \geq 1$ we have $y(t, s, \xi) \in \text{ int } D_0$. (iii) Finally, for the star-like closure of the set D_0 we have $D_0^* \subset D$.

In [19] we had a the following stability condition (Condition 1.5 of [19] with minor corrections added):

Condition D (i) For every $\xi \in D_0,\ t \geq s \geq 1\ \ y(t,s,\xi) \in D$ is defined for $1 \leq s \leq t < \infty$ and converges to x^* for $t \to \infty$ and we have with some $C_0, \alpha > 0$

$$\|\frac{\partial}{\partial \xi} y(t,s,\xi)\| \leq C_0 (s/t)^{\alpha}. \tag{3.22}$$

(ii) We assume that the initial condition ξ is in $\text{ int } D_{00}$, where $D_{00} \subset \text{ int } D_0$ is a compact domain which is invariant for (3.15) such that for any $t > s \geq 1$

$$y(t,s,D_{00}) = \{y(t,s,x) : x \in D_{00}\} \subset \text{ int } D_{00}.$$

Remark. The condition on the existence of D_0' can be removed if D itself is convex. Indeed, the ODE given by (3.15) becomes autonomous after a change of time scale $t = e^v$ (see below), thus part (i) of Condition D implies that the set

$$D_0'' = \{y : y = y(t,s,\xi), \xi \in D_0,\ t \geq s \geq 1\}$$

is invariant for the ODE. It is easy to see that it is also compact, so we can take for D_0' the convex envelope of D_0''. We will show below that part (ii) of Condition D follows from part (ii) of Condition 3.4. Finally, part (iii) of Condition 3.4 is a minor additional technical condition needed for the present paper.

We shall use subscripts to indicate partial derivatives below. Using a change of time-scale $t = e^v$, $s = e^u$, the inequality (3.21) is equivalent to the condition that for the solutions of the differential equation

$$\frac{d}{dv} z_v = G(z_v), \qquad z_u = \xi,\ u \geq 0,$$

denoted by $z(v,u,\xi)$ we have

$$\|z_\xi(v,u,\xi)\| \leq C_0 e^{-\alpha(u-v)}. \tag{3.23}$$

It can be shown that if for $\xi = x^*$ we can verify $\|z_\xi(v,u,x^*)\| \leq C_0' e^{-\alpha(u-v)}$ with some C_0' then (3.23) follows from the remaining components of Condition 3.4. Equivalently, it can be shown that if $\|y_\xi(t,s,x^*)\| \leq C_0'(s/t)^{\alpha}$ with some C_0' then (3.21) follows from the remaining components of Condition 3.4.

Setting

$$A^* = \frac{\partial G(x)}{\partial x}|_{x=x^*}, \tag{3.24}$$

we have $y_\xi(t,s,x^*) = e^{A^*(\log t - \log s)}$. The exponent α can be related to the eigenvalues of the Jacobian-matrix A^* as follows. Let

$$\alpha^* = \min_i \{-\Re \lambda_i(A^*)\}, \quad i = 1, ..., p, \tag{3.25}$$

where $\lambda_i(A^*)$ denote the eigenvalues of A^* and $\Re$ denotes real part. Then, denoting the spectral norm by $||.||_{sp}$ we have $||e^{A^*(\log t - \log s)}||_{sp} = e^{-\alpha^*(\log t - \log s)} = (s/t)^{\alpha^*}$. Since for any square matrix B we have $\lim_n ||B^n||^{1/n} = ||B||_{sp}$, we conclude that by taking

$$\alpha = \alpha^*_-, \tag{3.26}$$

where α^*_- denotes any number that is smaller than α^*, we have

$$||e^{A^*(\log t - \log s)}|| \leq C_0 e^{-\alpha(\log t - \log s)} = C_0 (s/t)^{\alpha} \tag{3.27}$$

with some $C_0 > 0$. If the Jordan-form of A^* is diagonal, then we can take $\alpha = \alpha^*$.

Lemma 3.1 *Condition 3.4 (ii) implies Condition D (ii).*

Proof: The proof is based on the observation that, introducing the notation $\Phi_v(\eta) = z(v, 0, \eta)$, for an arbitrary set of initial conditions $\Xi \subset D$ the set

$$D_{00} = \{z : z = \Phi_u(\eta),\ \eta \in \Xi,\ u \geq 0\}$$

is invariant for the associated ODE (3.15). Now let S denote the sphere in the z-space defined by $S = \{\zeta : |\zeta - x^*| = \delta > 0\}$ with some fixed small δ and let ζ_0 denote the point on S, where the trajectory $z(u, 0, \xi)$ hits S, say let $\zeta_0 = \Phi_{u_0}(\xi)$ with some $u_0 = u(\zeta_0)$. Now for any $\zeta \in S$ consider the inverse image of ζ under the flow Φ, i.e. consider the set of points

$$\Xi'(\zeta) = \{\eta' : \zeta = \Phi_u(\eta'),\ \text{for some } u \geq 0\}.$$

Let now $u(\zeta)$ be a continuously differentiable positive function on S, denoting the travel time from some initial point η up to ζ and define

$$\Xi = \{\eta : \zeta = \Phi_{u(\zeta)}(\eta), \zeta \in S\}.$$

Choosing $u(\zeta) \leq u(\zeta_0)$ for all ζ and ensuring that $u(\zeta)$ is very small whenever the angle between ζ and ζ_0 is larger than a fixed positive number the above defined set D_{00} will satisfy the second part of Condition D.

Finally, consider the perturbation term $\delta H(t, \omega)$. Following [19] and motivated by the application for the DFL scheme, we will use the following condition:

Condition 3.5 *$(\delta H(t, \omega))$ is a measurable M-bounded process, which is piecewise continuous in t almost surely, moreover there exists an $\varepsilon > 0$ such that for any fixed $q > 1$ and for any $s \geq 1$*

$$\sup_{s \leq \sigma \leq qs} \int_{\sigma}^{\tau(\sigma) \wedge q\sigma} \frac{1}{r} |\delta H(r, \omega)| dr = O_M(s^{-1/2-\varepsilon}). \tag{3.28}$$

It is no loss of generality to assume that $\varepsilon < 1/2$. We assume that the ε-s showing up here and in Condition 3.3 are identical.

Remark. In [19] the slightly weaker condition

$$\sup_{s \leq \sigma \leq qs} \int_{\sigma}^{\tau(\sigma) \wedge q\sigma} \frac{1}{r} |\delta H(r, \omega)| dr = O_M(s^{-1/2}) \tag{3.29}$$

was required (see Condition 1.6 of [19]). This is sufficient to establish a rate of convergence result for the moments.

The above condition seems to be hard to verify, since it involves $\tau(\sigma)$, which itself is defined in terms of the process (x_t). In fact, the condition seems to be artificially tuned so that the proof can be carried out. An alternative, seemingly more useful condition, implying Condition 3.5 would be

$$\sup_{s \leq \sigma \leq qs} \int_{\sigma}^{q\sigma} \frac{1}{r} |\delta H(r, \omega)| dr = O_M(s^{-1/2-\varepsilon}) \tag{3.30}$$

which is independent of the stopping time $\tau(\sigma)$). The latter is certainly satisfied if $\delta H(r, \omega) = O_M(r^{-1/2-\varepsilon})$. The prominent role of Condition 3.5 will become clear in the context of the DFL scheme, see (3.56) and Lemma 3.2. The following is given in Theorem 1.1 of [19]:

Theorem 3.1 *Consider the continuous-time recursive estimation process defined by (3.16) with the resetting mechanism (3.17) and (3.18). Assume that Conditions 3.1-3.5 are satisfied, moreover Condition 3.4 is satisfied with $\alpha > 1/2$. Then the solution (x_t) is defined for all $t \in [1, \infty)$ with probability 1 and $x_t = O_M(t^{-1/2})$. Moreover the following stronger result also holds: for any fixed $1 < q < \infty$ we have*

$$x_t^* = \sup_{t \leq s \leq qt} |x_s| = O_M(t^{-1/2}).$$

As has been noted in [19], end of Section 2, using the alternative resetting method (3.19) does not effect the validity of Theorem 1.1

Definition of $\overline{\alpha}$. In subsequent analysis a crucial role will be played by the *gap* between α, introduced in Condition 3.4 and $1/2$, therefore we introduce a separate notation: we write

$$\overline{\alpha} = \alpha - 1/2. \tag{3.31}$$

An example: a recursive estimation method is called a stochastic Newton method if the Jacobian-matrix of the right hand side of the associated ODE at $x = x^*$ is $-I$, where I is an identity matrix. Then we can take $\alpha^* = \alpha = 1$ and $\overline{\alpha} = 1/2$.

Let us now consider *discrete-time* processes of the form

$$x_{n+1} = x_n + \frac{1}{n+1}(H(n+1, x_n, \omega) + \delta H(n+1, \omega)), \qquad x_0 = \xi \in \text{ int } D_0. \tag{3.32}$$

Boundedness of the estimator sequence will be enforced by a *resetting* mechanism. Let $D_0 \subset D$ be a compact domain. If x_{n+1} leaves D_0 then we redefine x_{n+1} to be x_0. To formalize this: at any time n let x_{n+1-} denote the value of x computed at time $n+1$ by (3.32) and let

$$B_{n+1} = \{\omega : x_{n+1-} \notin \text{int} D_0\}. \tag{3.33}$$

Algorithm DR. A discrete-time recursive estimation process with resetting is defined as follows:

$$x_{n+1} = x_n + (1 - \chi_{B_{n+1}})\frac{1}{n+1}(H(n+1, x_n, \omega) + \delta H(n+1, \omega)) + \chi_{B_{n+1}}(x_0 - x_n). \tag{3.34}$$

Remark. Note that the correction term on the right hand side was $H(n, x_n, \omega)$ in [19]. The present notation fits the applications better: the estimator based on observations up to time n is updated by a new observation received at time $n+1$.

A standard way of analyzing this algorithm is to use continuous-time imbedding and this route has been followed in [19]. A more recent approach, in which the error that arises via this imbedding procedure is eliminated, is a discrete-time ODE method, developed in [25]. Here we follow the approach of [19], with a minor modifications. Let $(H^c(t, x, \omega))$ be the piecewise constant continuous-time extensions of $(H(n, x, \omega))$:

$$H^c(t, x, \omega) = H(n, x, \omega) \quad \text{for} \quad 1 \le n \le t < n+1. \tag{3.35}$$

Define $\delta H^c(t, x, \omega)$ in a similar manner. Let the exit time $\tau(\sigma)$ for any non-negative integer σ be defined as

$$\tau(\sigma) = \min\{n : n \text{ integer}, \ n > \sigma, x_{n-} \notin \text{int} D_0\}. \tag{3.36}$$

Condition 3.6 $(\delta H(n, \omega))$ *is a measurable* M*-bounded process, moreover there exists an* $\varepsilon > 0$ *such that for any fixed* $q > 1$ *and for any integers* $s \ge \sigma \ge 1$*, with* $[x]$ *denoting integer part, we have*

$$\sup_{s \le \sigma \le [qs]} \sum_{r=\sigma}^{\tau(\sigma) \wedge [q\sigma]} \frac{1}{r} |\delta H(r, \omega)| dr = O_M(s^{-1/2-\varepsilon}). \tag{3.37}$$

It is easy to see that in the course Condition 3.5 follows with the modified resetting mechanism (3.19). The following result is an easy corollary of Theorem 3.1 and has been established in [19] as Theorem 1.2:

Theorem 3.2 *Consider the discrete-time recursive estimation process with resetting defined by (3.34). Let* $(H^c(t, x, \omega))$ *be the piecewise constant continuous-time extensions of* $(H(n, x, \omega))$ *defined under (3.35). Assume that* $H^c(t, x, \omega)$ *satisfies Conditions 3.1- 3.4 and the latter condition is satisfied with* $\alpha > 1/2$*. Let* $\delta H^c(n, \omega)$ *satisfy Condition 3.6, with* $\tau(\sigma)$ *defined as in (3.36). Then we have* $x_n = O_M(n^{-1/2})$.

Let us now consider a general recursive estimation scheme developed in [11, 12, 51], see also [3, 13, 53], which will be called the Djereveckii-Fradkov-Ljung scheme, or the DFL scheme. Its basic building block is a parameter-dependent vector-valued process $(\overline{\phi}_n(x))$, with $x \in D \subset \mathbb{R}^p$, where D is an open domain, defined by the state-space equation

$$\overline{\phi}_{n+1}(x) = A(x)\overline{\phi}_n(x) + B(x)e_n, \tag{3.38}$$

with some non-random initial condition $\overline{\phi}_1(x)$, the value of which is often assumed to be zero. The dimensionality of $\overline{\phi}_n(x)$ will be denoted by r. In the analysis of [19], as in all other works on the analysis of the DFL scheme we have to ensure that for any choice of $x = x_n \in D$ the time-varying system

$$\phi_{n+1} = A(x_n)\phi_n + B(x_n)e_n, \quad \phi_0 = 0, \tag{3.39}$$

is bounded input-bounded output stable. This is ensured by the following condition:

Condition 3.7 *The functions $A(x), B(x)$ are three times continuously differentiable in D. Moreover, the family of matrices $A(x)$, $x \in D_0$, with D_0 being the pre-selected truncation domain, is jointly stable in the sense that there exist a single symmetric positive definite $r \times r$ matrix V and $0 < \lambda < 1$ such that for all $x \in D_0$*

$$A^T(x)VA(x) \leq \lambda V.$$

Discussion of the joint stability condition. In the case of recursive estimation of linear stochastic systems the joint stability condition can be satisfied by an appropriate realization of the state-system (3.38). Namely, in these cases (3.38) has the structure

$$\begin{aligned} \overline{\phi}_{1,n+1} &= A_1\overline{\phi}_{1,n} + B_1e_n, & (3.40)\\ \overline{\phi}_{2,n+1}(x) &= A_2(x)\overline{\phi}_{2,n}(x) + B_2(x)\overline{\phi}_{1,n+1}, & (3.41) \end{aligned}$$

where $\overline{\phi}_{1,n}$ is independent of x and is observable. Thus it is sufficient to ensure the joint stability of (3.41), which has an observable input. For any fixed x and non-singular $T = T(x)$ we have the system-equivalence

$$(\ A_2(x),\ B_2(x),\ I\ ) = (\ T(x)A_2(x)T^{-1}(x),\ T(x)B_2(x),\ T(x)^{-1}\ ), \tag{3.42}$$

and the latter realization can also be used to compute $\overline{\phi}_{2,n+1}(x)$. Assume that $A(x)$ is stable for all $x \in D$. Choosing $T(x)$ so that $T(x)A_2(x)T^{-1}(x)$ is a contraction for all $x \in D$ and assuming that $T(x)$ is continuous in x, it is easy to see that Condition 3.7 is satisfied for the transformed system with any compact $D_0 \subset D$. In addition, assuming that $(A_2(x), B_2(x), I)$ uniquely determines x, the same holds for the equivalent system $(\ T(x)A_2(x)T^{-1}(x),\ T(x)B_2(x),\ T(x)^{-1}\ )$.

Assuming joint stability of $(A(x))$ it follows that there exists some $c > 0$ such that for any sequence $(A(x_n))$ with $x_n \in D_0$ we have

$$||A(x_n)...A(x_0)|| \leq c\lambda^{n/2}. \tag{3.43}$$

In fact, this is the key property that we need in the analysis. An alternative method for ensuring the validity of (3.43) used in [3] is to require that the sequence $(A(x_n))$, or equivalently the sequence (x_n) is *slowly-varying*. This method will be discussed later.

The input noise (e_n) is assumed to satisfy two conditions (see Conditions 2.5, 2.6 of [19].)

Condition 3.8 *We assume that (e_n) is a wide-sense stationary process and that $|e_n|^2$ is in class M^*, i.e. for some $\varepsilon > 0$ we have*

$$\sup_n \mathrm{E}\ \exp \varepsilon|e_n|^2 < \infty.$$

Condition 3.8 is standard in the Chinese school for recursive estimation (see e.g. [7]) and is certainly satisfied for wide-sense stationary Gaussian sequences. The weaker condition that (e_n) is M-bounded is assumed also in the special case of (3.53), given as Example 1, p. 215 of [3], (see Condition (A'5) on p. 290 of [3]). The existence of finite moments of all orders for certain state-variables is required also in the general model of recursive estimation of [3], see Condition (A'5) on p. 290 of [3].

Discussion of Condition 3.8. Assume $\delta H(r,\omega) = 0$ identically and that no resetting takes place in the interval $[1,t]$. Then we have

$$x_t - y_t = \int_1^t \frac{1}{r}(H(r,x_r,\omega) - G(y_r))\ dr. \tag{3.44}$$

Now we can bound the right hand side form above in two ways as

$$\begin{aligned} |\int_1^t \frac{1}{r}(H(r,x_r,\omega) - G(x_r))\ dr| \quad &+ \quad \int_1^t \frac{1}{r}L|x_r - y_r|\ dr \\ |\int_1^t \frac{1}{r}(H(r,y_r,\omega) - G(y_r))\ dr| \quad &+ \quad \int_1^t \frac{1}{r}L_r|x_r - y_r|\ dr. \end{aligned} \tag{3.45}$$

In both cases we can apply the Bellman-Gronwall lemma. In the first case we need only the Lipschitz-continuity of G, while H may be even discontinuous, (which is the case e.g. for the signed LMS methods), but the first term is hard to analyze, unless H is a Markov-process for any fix x (see Chapter 1 of Part II of [3]). In the second case we need the Lipschitz-continuity of H and Condition 3.8 has to be imposed on L_r to ensure that the application of the Bellman-Gronwall lemma gives meaningful result. On the other hand the analysis of the first term is significantly simpler, since it is essentially the integral of a zero mean L-mixing process.

Condition 3.9 *We assume that (e_n) is L^+-mixing with respect to a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+)$.*

The role of this condition will be discussed in connection with Lemma 3.2, see also [19], Lemma 5.6 and Theorem 6.1. Now we are ready to define a random field $H(n,x,\omega)$ in terms of $\overline{\phi}_n(x)$ as follows:

$$H(n,x,\omega) = Q(\overline{\phi}_n(x)), \tag{3.46}$$

where for the sake of simplicity Q is a quadratic function from $\mathbb{R}^r$ to $\mathbb{R}^p$. An alternative, more general definition would be

$$H(n,x,\omega) = F(Q(\overline{\phi}_n(x)),x)), \tag{3.47}$$

where Q is quadratic, F is linear in Q and three times continuously differentiable in its second variable x. Also define the mean field

$$G(x) = \lim_{n\to\infty} \mathrm{E}\ Q(\overline{\phi}_n(x)). \tag{3.48}$$

It is easy to see that $G(x)$ is well-defined, since $\overline{\phi}_n(x)$ is asymptotically wide-sense stationary: in fact $\overline{\phi}_n(x) = \overline{\phi}_{*n}(x) + O_M(\beta^n)$, where $\overline{\phi}_{*n}(x)$ is wide-sense stationary and $0 < \beta < 1$ and thus

$$G(x) = \mathrm{E}\ Q(\overline{\phi}_n(x)) + O(\beta^n). \tag{3.49}$$

The estimation problem in the context of the DFL scheme is then to solve the non-linear algebraic equation

$$G(x) = 0$$

based on observations of $Q(\overline{\phi}_n(x))$. It is assumed that a unique solution x^* exists in D and in fact $x^* \in D_0$. In identification problems the estimation of x^* can be carried out in an off-line fashion, but this is not the case in stochastic adaptive control. Thus we focus on recursive estimation of x^*.

It is not difficult to see (cf. [19]) that under Conditions 3.7, 3.8 and 3.9 the piecewise constant continuous-time extension of the random field $H(n,x,\omega)$ defined by (3.46) satisfies Conditions

3.1, 3.2 and 3.3 with G defined under (3.48). In fact, in the latter condition $\delta G(t,x)$ decays exponentially fast to zero.

We use an iterative procedure, in which the estimate of x^* at time n will be denoted by x_n. To update this estimate we should use the correction term $Q(\overline{\phi}_n(x_n))$, but this frozen parameter value can not be easily computed. In fact in stochastic adaptive control problems it can not be computed at all. Hence we generate an on-line approximation of $Q(\overline{\phi}_n(x_n))$ and thus we arrive at the following first version of the DFL method: define recursively

$$\phi_{n+1} = A(x_n)\phi_n + B(x_n)e_n \tag{3.50}$$

$$x_{n+1} = x_n + \frac{1}{n}Q(\phi_{n+1}) \tag{3.51}$$

with initial conditions $x_0 = \xi \in \text{int}D_0$ and ϕ_0 a constant, non-random initial state. It is assumed that $Q(\phi_{n+1})$ is *computable* by coupling a physical system with our computer.

Discussion on the DFL scheme. The applicability of this general estimation scheme in the theory of recursive identification of linear stochastic systems has been discussed in much details [53], albeit its analysis has not been complete. Further examples of application are given in [3]. Here also a rigorous and detailed analysis of a non-linear modification of the DFL scheme is given, using a Markovian dynamics in generating the state sequence (ϕ_n). This setup extends the range of applicability of the method, but the verification of the existence of the solution of a Poisson-equation, (see Condition (A.4) of Chapter 1.1, Part II in [3]), seems to be hard. A special case of the DFL scheme is stochastic linear regression, in which $\overline{\phi}_n$ does not depend on x at all and $H(n,x,\omega)$ is of the form

$$H(n,x,\omega) = Q(l(\overline{\phi}_n, x)) \tag{3.52}$$

with l being linear both in $\overline{\phi}$ and x, has been analyzed in [7, 14, 47].

It is well-known from simulations that the DFL scheme may diverge, unless some precaution is taken. The above procedure will therefore be modified so that the estimates x_n will be enforced to stay in a compact domain $D_0 \subset D$, such that $x^* \in \text{int}D_0$. This will be achieved by a *resetting* mechanism: if x_{n+1} leaves D_0 we redefine it to be x_0. To formalize this procedure let x_{n+1-} denote the value of x computed at time $n+1$ by (3.51). Then if $x_{n+1-} \notin \text{int}D_0$ then we reset it to its initial value ξ. To formalize the procedure let

$$B_{n+1} = \{\omega : x_{n+1-} \notin \text{int}D_0\}.$$

Then we define:

Algorithm DFL: The Djereveckii-Fradkov-Ljung or DFL scheme with resetting:

$$\phi_{n+1} = A(x_n)\phi_n + B(x_n)e_n \tag{3.53}$$

$$x_{n+1} = x_n + (1 - \chi_{B_{n+1}})\frac{1}{n+1}Q(\phi_{n+1}) + \chi_{B_{n+1}}(x_0 - x_n). \tag{3.54}$$

An additional stopping time is used in [3] to ensure the validity of (3.43) by ensuring that the sequence $(A(x_n))$, or equivalently the sequence (x_n) is *slowly-varying*. Following [3], (3.1.2) on p. 291, for any positive integer σ define the stopping time

$$\nu(\sigma) = \min\{n : n \text{ integer}, \ n > \sigma, |x_n - x_{n-1}| > \delta\}, \tag{3.55}$$

where σ is some fixed positive number. It is well-known that if δ is sufficiently small, then (3.43) holds. However, the a priori determination of a right value of δ seems to be hard.

Discussion of the "boundedness condition". The eventual divergence of the DFL scheme is often dealt with the controversial "boundedness condition" first formulated in [51], requiring that the estimator process visits a compact domain of attraction of the ODE infinitely often. A lucid exposition of the underlying principle is given in [52] see Lemma 1.12, which is considered there as the key tool for the ODE method. Almost sure convergence using the above "boundedness condition" has also been established for a non-linear, Markovian extension of the DFL method

in [3], Part II, Chapter 1.9, Theorem 15. Unfortunately, the "boundedness condition" is much too restrictive: it is a condition on the process itself that we analyze and it is not clear at all if it is satisfied even for basic methods such as RPE for ARMA-processes.

One way to enforce the boundedness of the estimation process is to consider a compact truncation domain containing the true parameter in its interior and to "project" the estimator back to this domain if it would leave it, see [51, 45]. It is easy to see that this procedure may fail even for deterministic algorithms, namely the ODE which approximates the evolution of the discrete time algorithm may force us to move out of the truncation domain. A sophisticated extension of the projection method using *expand in g truncations* has been developed by H.F. Chen.

A rigorous treatment of the boundedness problem has been given in [3], where the estimator process is stopped if it leaves a prescribed compact domain containing the true parameter in its interior. Denoting by $\Omega' \subset \Omega$ the event that the estimator process is never stopped, the almost sure convergence of the estimator process has been established on Ω', see [3], Part II, Chapter 1.6, Proposition 11. But convergence with probability strictly smaller than 1 is not satisfactory from the practical point of view. The above *truncated* version of the DFL-methods has been given and analyzed in [19].

The definition of the truncation domain requires some a priori knowledge of the system parameters no matter what truncation procedure we use. This may seem to be a restrictive assumption but even deterministic iterative methods for optimization may fail without good initialization.

In practice we start with an initial value and a truncation domain which may or may not satisfy our conditions. If it does not and the solution trajectory of the associated ODE starting at $x_0 = \xi$ does hit the boundary of D_0, then a heuristic argument, following [19], shows that the estimator process will be likely to hit the neighborhood of the same point of the boundary of the truncation domain. This phenomenon can be detected during the computations and a larger truncation domain can be chosen. Such an adaptive choice of the truncation domain has not yet been studied. A special case when the boundedness problem does not arise is the use of a stochastic regression approach, such as extended least squares (ELS), see [53].

To connect the DFL scheme with Algorithm DR define

$$\delta H(n,\omega) = Q(\phi_n) - Q(\bar{\phi}_n(x_n)). \tag{3.56}$$

Then (3.54) can be written in the form of (3.34). A critical point in the analysis of the DFL scheme is that the perturbation term $\delta H(n,\omega)$ is *not given a priori*, rather it is defined via the recursive procedure itself. In fact, the analysis of $\delta H(n,\omega)$ is a substantial component of the convergence analysis of the DFL-method, which has been worked out in [19], Section 5 and 6, leading to the following result (cf. Lemma 5.6 of [19]):

Lemma 3.2 *Consider the DFL scheme defined by (3.53)-(3.54). Assume that Conditions 3.7, 3.8, 3.9 are satisfied. In addition assume that Condition 3.4 is satisfied with $\alpha > 1/2$. Then $(\delta H(n,\omega))$ defined by (3.56) is an M-bounded process, moreover there exists an ε with $0 < \varepsilon < 1/2$ such that for any fixed $q > 1$ and for any integer $s \geq 1$ and integers σ*

$$\sup_{s \leq \sigma \leq [qs]} \sum_{\sigma}^{\tau(\sigma) \wedge [q\sigma]} \frac{1}{r} |\delta H(r,\omega)| dr = O_M(s^{-1/2-\varepsilon}). \tag{3.57}$$

In short: $(\delta H(n,\omega))$ satisfies Condition 3.6. Postulating the validity of Condition 3.4 we conclude that all conditions of Theorem 3.2 are satisfied and thus we get:

Theorem 3.3 *Consider the DFL scheme defined by (3.53)-(3.54). Assume that Conditions 3.7, 3.8 and 3.9 are satisfied. In addition assume that Condition 3.4 is satisfied with $\alpha > 1/2$. Then we have $x_n = O_M(n^{-1/2})$.*

Discussion of the result. A special feature of the above result is that the *moments* of the estimation error are bounded from above. The only alternative result on the moments of the estimation error in the context of the DFL scheme seems to be Proposition 24 of [3], Part II,

Chapter 1.10, where the L_2 moments of the error of the stopped process is shown to be of the order $1/n$.

Almost sure convergence of the DFL scheme has been stated in [51] using the controversial "boundedness condition", requiring that the estimator process visits a compact domain of attraction of the ODE infinitely often. See also [52], Lemma 1.12 for a related result. Almost sure convergence using the above "boundedness condition" has also been established for a non-linear, Markovian extension of the DFL method in [3], Part II, Chapter 1.8, Theorem 15. The almost sure convergence of the estimator process has been established on a set $\Omega' \subset \Omega$ of probability strictly less than 1, see [3], Part II, Chapter 1.6, Proposition 11 and Chapter 3.4, Theorem 17.

An alternative set of results are obtained for stochastic regression models developed in [47]. Results on the rate of almost sure convergence are given in [7] and [14]. See also Theorem 2 of [4] or Theorem 1 of [10]. The main shortcoming of stochastic regression, such as extended least squares (ELS), see [53], compared to the DFL scheme is that its range of applicability is limited. E.g in estimating an ARMA process by ELS we must impose the condition that the polynomial $C - 1/2$ is positive real.

Further discussion on mixing conditions. L-mixing and ϕ-mixing can both be used to derive two main results of [19], (Theorems 1.1 and 1.2), restated here as Theorem 3.1 and Theorem 3.2. In both results the key technical device is an improved Hölder-inequality, see Lemma 8.1 of the Appendix, or Chapter 7.2 of [15], or Appendix III of [33]. An improved Hölder-inequality of [15] is restated as Lemma 8.2. The situation is quite different for the DFL scheme, where L^+-mixing has been heavily exploited for deriving Theorem 3.3, in particular in proving Lemma 3.2 (see Sections 5 and 6 of [19], in particular Theorem 6.1 in [19]).

4 Strong approximation of the estimation error

The main result of the present paper is a significant extension of Theorem 2.4 for the three, closely related recursive estimation schemes presented in the previous section. These extensions will be stated and proved in this section. The analysis will be carried out in detail for Algorithm CR, given by (3.16) and the resetting mechanism (3.17) and (3.18), see Theorem 4.1. The proof is non-trivial and relies on the results of [17, 19] and [21]. The corresponding results for Algorithm DR and Algorithm DFL will then follow by relatively simple arguments. The extension of the two other main results for the RPE method, given in Section 2 as Theorems 2.5 and 2.6, will be given in the next two sections. Note that the conditions for the next theorem are identical with the conditions of Theorem 3.1.

Theorem 4.1 *Consider the continuous-time recursive estimation scheme Algorithm CR given by (3.16) with the resetting mechanism (3.17) and (3.18). Assume that Conditions 3.1-3.5 are satisfied and Condition 3.4 is satisfied with $\alpha > 1/2$. Then the solution of (3.16), (x_t), is defined for all $t \in [1, \infty)$ with probability 1 and we have with*

$$\varepsilon_x = \min(\overline{\alpha}, \varepsilon)_-,$$

where c_- is any number smaller than c, $\overline{\alpha}$ is given by (3.31) and ε is given in Condition 3.5,

$$x_t - x^* = \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H(s, x^*, \omega) ds + O_M(t^{-1/2-\varepsilon_x}). \tag{4.1}$$

Discussion of the result. The bound $O_M(t^{-1/2-\varepsilon_x})$ can not be improved in general. Indeed, let $\delta H(t, \omega) = 0$, then $\varepsilon_x = \overline{\alpha}_- = \alpha_- - 1/2$, where $\alpha = \alpha^*_-$ (see (3.31), (3.26) and (3.25)). Thus

$$-1/2 - \varepsilon_x = -\alpha^*_-.$$

Consider now a linear process with additive, state-independent, bounded noise, i.e. let $H(t, x, \omega) = A^* x + u_t$, where (u_t) is a zero-mean L-mixing bounded process. Then Algorithm CR reads

$$\dot{x}_t = \frac{1}{t}(A^* x_t + u_t), \qquad x_1 = \xi. \tag{4.2}$$

Assuming that A^* is stable, the boundedness of (u_t) implies the boundedness of (x_t), hence taking a sufficiently large truncation domain no resetting will take place ever. Obviously we have $x^* = 0$ and we can write the exact equality

$$x_t = \left(\frac{\partial}{\partial \xi} y(t, 1, x^*)\right) \cdot \xi + \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H(s, x^*, \omega) ds \tag{4.3}$$

with

$$\frac{\partial}{\partial \xi} y(t, s, x^*) = e^{A^*(\log t - \log s)}.$$

Thus the residual term, the first term on the right hand side of (4.3), is $e^{A^* \log t}\xi$, since now $s = 1$. Thus we have

$$||e^{A^* \log t}||_{sp} = t^{-\alpha^*}, \tag{4.4}$$

and since for any square matrix B we have $||B|| \geq ||B||_{sp}$, we conclude that

$$||e^{A^* \log t}|| \geq t^{-\alpha^*}. \tag{4.5}$$

Thus there exists a ξ such that

$$|e^{A^* \log t}\xi| \geq t^{-\alpha^*}|\xi|, \tag{4.6}$$

implying that the result of the theorem is sharp.

To interpret this result note that the matrix $(\frac{\partial}{\partial \xi})y(t, s, x^*)$ is the sensitivity matrix, which indicates the relative effect of a perturbation of the initial condition at time s on the solution of (3.15) at time t. Thus the dominant term on the right hand side represents the cumulative effect of the ideal correction terms $\frac{1}{s}H(s, x^*, \omega)$ at time t. A similar representation of the error $x_t - x^*$ for classical Robbins-Monroe processes, had been implicitly used already in [54]. The above dominant term has been explicitly presented for a class of stopped stochastic approximation processes in Lemma 3.1 of [65].

The novelty of the present result is that it is stated for a general recursive estimation scheme, that can handle the widely-used DFL scheme, a crucial boundedness assumption enforced by a resetting mechanism and a tight upper bound for the residual term has been obtained. A relatively straightforward corollary of Theorem 4.1 is the following discrete-time result, in which the conditions are identical with the conditions of Theorem 3.2:

Theorem 4.2 *Consider the discrete-time recursive estimation process Algorithm DR with resetting defined by (3.34). Let $(H^c(t, x, \omega))$ be the piecewise constant continuous-time extension of $(H(n, x, \omega))$ defined under (3.35). Assume that $(H^c(t, x, \omega))$ satisfies Conditions 3.1-3.4 and Condition 3.4 is satisfied with $\alpha > 1/2$. Let $\delta H(n, \omega)$ satisfy Condition 3.6, with $\tau(\sigma)$ defined in (3.36). Then we have, with $\varepsilon_x = \min(\overline{\alpha}, \varepsilon)_-$, where $\overline{\alpha}$ is given by (3.31) and ε is given by Condition 3.6,*

$$x_N - x^* = \sum_{n=1}^{N} \frac{\partial y}{\partial \xi}(N, n, x^*) \frac{1}{n} H(n, x^*, \omega) + O_M(N^{-1/2-\varepsilon_x}).$$

Specializing the last result to the DFL scheme we get a result that is very useful for applications (see Section 7):

Theorem 4.3 *Consider the DFL scheme defined by (3.53)-(3.54). Assume that the state-space equation (3.38) satisfies Condition 3.7, the noise process (e_n) satisfies Condition 3.8 and 3.9 and the associated ODE satisfies Condition 3.4 with $\alpha > 1/2$. Let $\varepsilon_x = \min(\overline{\alpha}, \varepsilon)_-$, where $\overline{\alpha}$ is defined under (3.31) and ε is given by Lemma 3.2. Then we have*

$$x_N - x^* = \sum_{n=1}^{N} \frac{\partial y}{\partial \xi}(N, n, x^*) \frac{1}{n} Q(\overline{\phi}_n(x^*)) + O_M(N^{-1/2-\varepsilon_x}).$$

Remark: The proof of Lemma 5.6 in [19], based on Theorem 6.1 of the same paper, implies that in Condition 3.5 we have $\varepsilon < 1/2$. Thus in the present case it is not our choice to have $\varepsilon < 1/2$. It follows that the upper bound for the residual term can not be as small as $O_M(N^{-1})$, in contrast to what we had for the off-line prediction error method for ARMA-processes, see Theorem 2.2.

The above results take a particularly attractive form for partially *stochastic Newton methods*. A recursive estimation method is called a partially stochastic Newton method if the Jacobian-matrix of the right hand side of the associated ODE at $x = x^*$ is of the form

$$\begin{pmatrix} -I & 0 \\ X & Y \end{pmatrix},$$

where I is an identity matrix. An example: the standard recursive prediction error estimation of ARMA processes, in which both the system-parameter θ^* and the Hessian of asymptotic cost-function R^* are estimated and the estimates of the system-parameters are updated using Newton-like steps, is a partially stochastic Newton method with respect to the system-parameters.

The above decomposition of the Jacobian is in one-to-one correspondence with the splitting of the parameter-vector x as $x = (x^1, x^2)$. With this notation it is easy to see that

$$\frac{\partial}{\partial \xi^1} y(t, s, \xi)_{|\xi = x^*} = \left(\frac{s}{t} I, 0 \right)$$

for $s \leq t$ and the statement of Theorem 4.3 simplifies to the following:

Theorem 4.4 *Assume that the conditions of Theorem are satisfied and that we can split the parameter-vector x as $x = (x^1, x^2)$ so that the estimation method is a partially stochastic Newton method with respect to x^1. Let (Q^1, Q^2) be the corresponding splitting of Q. Then we have with the same ε_x as in Theorem 4.3*

$$x^1_N - x^{1*} = \frac{1}{N} \sum_{n=1}^{N} Q^1(\overline{\phi}_n(x^*)) + O_M(N^{-1/2-\varepsilon_x}).$$

Theorem 4.4 is an extension of Theorem 2.4 to general partially stochastic Newton methods, but with a weaker error term, since $\varepsilon_x < 1/2$.

The result given as (2.23) can also be extended. Let the off-line estimator $\hat{x}_N$ of x^* be defined as the solution of

$$U_N(x) = \sum_{n=1}^{N} Q^1(\overline{\phi}_n(x)) = 0$$

with respect to x. The handling of multiple solutions is precisely described in [18]. Then it is easy to see that Theorem 2.2 can be extended andnoting that the Jacobian-matrix of the right hand side of the associated ODE at $x = x^*$ is of the form given above, we get for the first component of $\hat{x}$

$$\hat{x}^1_N - x^{1*} = \frac{1}{N} \sum_{n=1}^{N} Q^1(\overline{\phi}_n(x^*)) + O_M(N^{-1}).$$

Combining this with Theorem 4.4 and writing $\widehat{\widehat{x}}_N = x_N$ we get

$$\widehat{\widehat{x}}^1_N - \hat{x}^1_N = O_M(N^{-1/2-\varepsilon_x}) \tag{4.7}$$

which is an extension of (2.23), albeit with a weaker error term.

Discussion of the result. To *compare* these results with results of [3] and [45] we note that the limit results of [3] are of classical nature: weak convergence and CLT (central limit theorem), which are not strong enough for calculating performance degradation that we called pathwise cumulative regret. The same remark applies to the weak-convergence results of [45].

In the case of stochastic regression methods, developed in [47] andextended in [7] and [14], tight bounds for the almost sure rate of convergence of the estimator process are given. But

even these results are not applicable in general to get exact asymptotic results for the pathwise cumulative regret, except in very special cases, such as the minimum-variance self-tuning regulator for ARX-systems, see [49]. For ARMAX-systems these techniques yield only qualitative results, see [48].

Further discussion on mixing conditions. The proof of Theorem 4.1 relies on a moment inequality for weighted multiple integrals of L-mixing processes given in [21]. It is likely that this result can be extended to ϕ-mixing processes, since it is based on the repeated use of an improved Hölder-inequality, which does have its variant for ϕ-mixing processes, see Lemmas 8.1 and 8.2 of the Appendix andChapter 7.2 of [15] for further results. Thus it is likely that L-mixing and ϕ-mixing can both be used to derive the results of the present section for Algorithm CR and Algorithm DR, given as Theorems 4.1 and 4.2.

The situation is quite different for the DFL scheme, where L^+-mixing has already been heavily exploited for getting the rate of convergence of higher order moments, see Theorem 3.3. Furthermore, L^+-mixing is very much used in the context of all three algorithms (Algorithm CR, Algorithm DR and the DFL scheme) in deriving the results of Sections 4 and 5. Moreover, the formulation of the main result of Section 5 is given in terms of the concept of L-mixing. It is not clear if a similar result holds in the context of ϕ-mixing. Even the following simple related problem seems to be open: under what conditions is the response of an exponentially stable linear filter, with a ϕ_p-mixing process as its input, ϕ_p-mixing ?

Proof of Theorem 4.1: Assume $x^* = 0$. Also we can assume that $\delta G = 0$, namely the term $\delta G(t, x_t)$ can be merged with $\delta H(t, \omega)$. Indeed, the condition that $\delta G(t, x) = O(t^{-1/2-\varepsilon})$ uniformly in x for $x \in D_0$, see Condition 3.3, implies that Condition 3.5 remains valid when $\delta H(t, \omega)$ is replaced by $\delta H(t, \omega) + \delta G(t, x_t)$.

Let us consider the process (x_t) on the interval $[s, qs)$ with $s \geq 1, \quad q > 1$ and let $\overline{y}_t$ denote the solution of the ordinary differential equation (3.15) starting from x_s at time s. Let C_s denote the event that x_t hits ∂D_0 in $[s, qs)$. Then we can write

$$x_t - \overline{y}_t = (1 - \chi_{C_s}) \int_s^t \frac{\partial}{\partial \xi} y(t, r, x_r) \cdot \frac{1}{r} \left(\overline{H}(r, x_r, \omega) + \delta H(r, \omega) \right) dr + \chi_{C_s}(x_t - \overline{y}_t) \qquad (4.8)$$

with $\overline{H}(r, x, \omega) = H(x, r, \omega) - G(r, x)$ by using Lemma 8.6 of the Appendix. Let us now take into account the fact that $y_\xi(t, r, x)$ and $\overline{H}(r, x, \omega)$ are continuously differentiable with respect to x. Hence we can write

$$\frac{\partial}{\partial \xi} y(t, r, x_r) = \frac{\partial}{\partial \xi} y(t, r, 0) + \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t, r, \lambda x_r) d\lambda \cdot x_r$$

and

$$\overline{H}(r, x_r, \omega) = \overline{H}(r, 0, \omega) + \int_0^1 \frac{\partial}{\partial x} \overline{H}(r, \lambda x_r, \omega) d\lambda \cdot x_r.$$

Substituting into (4.8) we get that the first integral on the right hand side of (4.8) can be written as the sum of the following five terms:

$$\begin{aligned}
I_1 &= \int_s^t \frac{\partial}{\partial \xi} y(t, r, 0) \cdot \frac{1}{r} \overline{H}(r, 0, \omega) dr \\
I_2 &= \int_s^t \frac{\partial}{\partial \xi} y(t, r, 0) \cdot \frac{1}{r} \int_0^1 \frac{\partial}{\partial x} \overline{H}(r, \lambda x_r, \omega) d\lambda \cdot x_r dr \\
I_3 &= \int_s^t \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t, r, \lambda x_r) d\lambda \cdot x_r \cdot \frac{1}{r} \overline{H}(r, 0, \omega) dr \\
I_4 &= \int_s^t \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t, r, \lambda x_r) d\lambda \cdot x_r \cdot \frac{1}{r} \int_0^1 \frac{\partial}{\partial x} \overline{H}(r, \lambda' x_r, \omega) d\lambda' \cdot x_r dr \\
I_5 &= \int_s^t \frac{\partial}{\partial \xi} y(t, r, x_r) \cdot \frac{1}{r} \delta H(r, \omega) dr.
\end{aligned}$$

We will later also write $I_1 = I_{1,t} = I_{1,t,s}$ when we want to emphasize the dependence of I_1 on t and s. Then we can write

$$x_t - \overline{y}_t = (1 - \chi_{C_s})(I_1 + I_2 + I_3 + I_4 + I_5) + \chi_{C_s}(x_t - \overline{y}_t). \tag{4.9}$$

We will approximate I_2 and I_3 so that we replace λx_r and $\lambda' x_r$ by 0 and define

$$\begin{aligned} I_2^* &= \int_s^t \frac{\partial}{\partial \xi} y(t,r,0) \cdot \frac{1}{r} \int_0^1 \frac{\partial}{\partial x} \overline{H}(r,0,\omega) d\lambda \cdot x_r dr \\ I_3^* &= \int_s^t \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t,r,0) d\lambda \cdot x_r \cdot \frac{1}{r} \overline{H}(t,0,\omega) dr. \end{aligned}$$

For the sake of notational homogenity we will also write $I_1 = I_1^*$.

Lemma 4.1 *We have for fixed q and any $s \le t \le qs$,*

$$x_t - \overline{y}_t = I_1^* + I_2^* + I_3^* + O_M(s^{-1/2-\varepsilon}). \tag{4.10}$$

Remark: It ill be clear from proof that in the case $\delta H(t,\omega) = 0$ the last term becomes $O_M(s^{-1})$. Indeed the error term $O_M(s^{-1/2-\varepsilon})$ shows up only in the last step of the proof, in the estimation of the effect of I_5. Thus a key factor in the accuracy of the ODE approximation is the perturbation term $\delta H(t,\omega)$.

Proof: *Estimation of I_2.* We claim that for $s \le t \le qs$ we have

$$I_2 = I_2^* + O_M(s^{-1}). \tag{4.11}$$

Indeed, fix λ and integrate first with respect to r. We expand $\frac{\partial}{\partial x}\overline{H}^i(r, \lambda x_r, \omega)$ (for $i = 1, \dots, p$) into a Taylor-series about 0 once more to obtain:

$$\frac{\partial}{\partial x}\overline{H}^i(r,\lambda x_r,\omega) = \frac{\partial}{\partial x}\overline{H}^i(r,0,\omega) + \left(\int_0^1 \frac{\partial^2}{\partial x^2}\overline{H}^i(r,\lambda'\lambda x_r,\omega) d\lambda' \right) \cdot x_r.$$

The expression under the integral term here can be shown to be $O_M(1)$ by the same argument that we used above, since $\overline{H}$ is assumed to have continuous third derivatives almost surely which are also M-bounded. Thus we get $\frac{\partial}{\partial x}\overline{H}(r,\lambda x_r,\omega) = \frac{\partial}{\partial x}\overline{H}(r,0,\omega) + O_M(r^{-1/2})$. Integration with respect to λ from 0 to 1 and multiplication by $r^{-1}x_r = O_M(r^{-3/2})$ yields an error term $O_M(r^{-2})$. Finally since $\|\frac{\partial}{\partial \xi} y(t,r,0)\| \le C_0(r/t)^\alpha$ we get

$$I_2 = \int_s^t \frac{\partial}{\partial \xi} y(t,r,0) \cdot \frac{1}{r} \frac{\partial}{\partial x} \overline{H}(r,0,\omega) \cdot x_r dr + O_M(s^{-1}) \tag{4.12}$$

as stated. Note that the dominant term can be estimated by using the moment inequality given as Theorem 8.1. Thus we also get $I_2 = O_M(s^{-1/2})$.
Estimation of I_3. We claim that for $s \le t \le qs$ we have

$$I_3 = I_3^* + O_M(s^{-1}). \tag{4.13}$$

Indeed, in the inner integrand of I_3 we can write

$$\frac{\partial^2}{\partial \xi^2} y(t,r,\lambda x_r) = \frac{\partial^2}{\partial \xi^2} y(t,r,0) + \left(\int_0^1 \frac{\partial^2}{\partial \xi^3} y(t,r,\lambda'\lambda x_r) d\lambda' \right) \cdot x_r, \tag{4.14}$$

where the last term is to be interpreted as the product of a 4-tensor with a 1-tensor yielding a 3-tensor, thus interpreting $\cdot$ as a tensor product. Substituting (4.14) into the expression of I_3 we get for fixed λ, λ' the product of the following two terms:

$$\begin{aligned} &\frac{\partial^2}{\partial \xi^2} y(t,r,0) \cdot x_r \cdot \frac{1}{r} \overline{H}(r,0,\omega) = O_M(r^{-3/2}) \\ &\frac{\partial^3}{\partial \xi^3} y(t,r,\lambda'\lambda x_r) \cdot x_r \cdot x_r \cdot \frac{1}{r} \overline{H}(r,0,\omega) = O_M(r^{-2}), \end{aligned}$$

where we used the fact that the partial derivatives of $y(t,r,\xi)$)with respect to ξ are bounded by a deterministic constant, see Lemma 8.8 of the Appendix. Integrating from s to t the contribution of the integral of the second term is $O_M(s^{-1})$, thus we get

$$I_3 = \int_s^t \left(\frac{\partial^2}{\partial \xi^2} y(t,r,0) \cdot x_r \right) \cdot \frac{1}{r} \overline{H}(r,0,\omega) dr + O_M(s^{-1}). \tag{4.15}$$

as stated. Note that the expected upper bound $I_3 = O_M(s^{-1/2})$ can not be readily derived from the above approximation: we can not use the moment inequality given as Theorem 8.1 since the weights x_r are random!

Estimation of I_4. We claim that for $s \le t \le qs$ we have

$$I_4 = O_M(t^{-1}). \tag{4.16}$$

Indeed, by Theorem 3.1 we have $x_r = O_M(r^{-1/2})$, hence for fixed λ, λ' the contribution of the term $r^{-1} x_r \cdot x_r$, interpreted as an appropriate tensor product, is $O_M(r^{-2})$. On the other hand $\|y_{\xi\xi}(t,r,x)\| \le C_0'(r/t)^\alpha \le C_1'$ with some $C_0', C_1' > 0$, uniformly in x for $x \in D_0$, cf. Lemma 8.8. Thirdly

$$\|\frac{\partial}{\partial x} H(r, \lambda' x_r, \omega)\| \le \sup_{x \in D_0^*} \|\frac{\partial}{\partial x} H(r,x,\omega)\| \triangleq H_x^*(x,r,\omega), \tag{4.17}$$

where D_0^* denotes the star-like closure of D_0. Since by assumption $D_0^* \subset \mathrm{int} D$ and the partial derivative $H_{xx}(r,x,\omega)$ exists and is continuous almost surely and is M-bounded, we get by the maximal inequality given as Theorem 8.3 in the Appendix that the right hand side of (4.17) is $O_M(1)$. Hence we finally get, using the triangle inequality, that

$$I_4 = O_M\left(\int_s^t C_0 (r/t)^\alpha r^{-2} dr \right) = O_M(s^{-1})$$

as stated.

Estimation of the effect of $I_{5,t}$. We claim that for $s \le t \le qs$

$$(1 - \chi_{C_s})\ I_{5,t} = O_M(s^{-1/2-\varepsilon}). \tag{4.18}$$

Indeed, we have

$$(1 - \chi_{C_s})\ |I_{5,t}| \le (1 - \chi_{C_s}) \int_s^{t \wedge \tau(s)} ||\frac{\partial}{\partial \xi} y(t,r,0)||\ \frac{1}{r} |\delta H(r,\omega)|\ dr$$

since for $t > \tau(s)$ we have $1 - \chi_{C_s} = 0$. Noting that $\|y_\xi(t,r,0)\| \le C_0$ and taking into account Condition 3.5 we get the claim.

Write now

$$x_t - \overline{y}_t = I_1 + I_2 + I_3 + I_4 + (1 - \chi_{C_s}) I_5 - \chi_{C_s}(I_1 + I_2 + I_3 + I_4) + \chi_{C_s}(x_t - \overline{y}_t), \tag{4.19}$$

and estimate the contribution of the last two terms. Note that

$$I_1 + I_2 + I_3 + I_4 = \int_s^t \frac{\partial}{\partial \xi} y(t,r,x_r) \cdot \frac{1}{r} \overline{H}(r, x_r, \omega)\ dr = O_M(1). \tag{4.20}$$

Indeed, $||y_\xi(t,r,x_r)|| \le C_0$ and $\overline{H}(r,x_r,\omega)$ is M-bounded, see the argument leading to (4.17). Similarly $|x_t - \overline{y}_t| = O_M(1)$, actually we have $|x_t - \overline{y}_t| = O(1)$. As for χ_{C_s} we have the following lemma that has been given as Lemma 2.3 in [19].

Lemma 4.2 *Consider the continuous-time recursive estimation scheme given by (3.16) with the resetting mechanism (3.17) and (3.18). Assume that Conditions 3.1-3.5 are satisfied. Let* C_s *denote the event that* x_t *hits* ∂D_0 *in the interval* $[s, qs)$. *Then for any* $m \ge 1$ *we have* $P(C_s) = O(s^{-m})$.

Thus the contribution of the last two terms in (4.19) is $O_M(s^{-m})$ for any $m \geq 1$ and with this Lemma 4.1 has been proved.

Our next step is to show that the dominant term is I_1^*, i.e. the terms I_2^* and I_3^* are negligible. This is stated in the next lemma, which is the *key lemma* for the proof of Theorem 4.1. Its proof requires a new tool, specially designed for the present application: moment inequalities for double integrals of L-mixing processes.

Lemma 4.3 *We have for* $s \leq t \leq qs$

$$x_t - \overline{y}_t = I_1^* + O_M(s^{-1/2-\varepsilon}).$$

Proof: In order to obtain sharper estimates of I_2^* and I_3^* let us write

$$I_2^* + I_3^* = \int_s^t g_r x_r \, dr, \tag{4.21}$$

where the matrix-valued process (g_r) is defined by

$$g_r = \frac{\partial}{\partial \xi} y(t,r,0) \cdot \frac{1}{r} \frac{\partial}{\partial x} \overline{H}(r,0,\omega) + \frac{\partial^2}{\partial \xi^2} y(t,r,0) \cdot \frac{1}{r} \overline{H}(r,0,\omega). \tag{4.22}$$

Thus we can write Lemma 4.1 as

$$x_t = \overline{y}_t + I_1^* + \int_s^t g_r x_r dr + O_M(s^{-1/2-\varepsilon}). \tag{4.23}$$

If we had $x_r = x$ a small constant, then we could write the integral on the right hand side of (4.23) as $\int_s^t g_r dr \cdot x$, which then could be estimated by the moment inequality give as Theorem 8.1, since both $\overline{H}_x(r,0,\omega)$ and $\overline{H}(r,0,\omega)$ are zero-mean L-mixing processes. If x is small then the contribution of this term will be negligable. To show that the second term in (4.23) is indeed negligible we iterate (4.23), i.e. substitute x_r by the expression that is given by (4.23). Writing $I_1^* = I_{1,t}^*$ we get

$$x_t = \overline{y}_t + I_{1,t}^* + \int_s^t g_r \left(\overline{y}_r + I_{1,r}^* + \int_s^r g_p x_p dp + O_M(s^{-1/2-\varepsilon}) \right) dr + O_M(s^{-1/2-\varepsilon}). \tag{4.24}$$

Let us set

$$J_1 = \int_s^t g_r \overline{y}_r dr, \qquad J_2 = \int_s^t g_r I_{1,r}^* dr, \qquad J_3 = \int_s^t g_r \int_s^r g_p x_p \, dp \, dr.$$

The last term of the double integral in (4.24) yields $\int_s^t g_r O_M(s^{-1/2-\varepsilon}) dr = O_M(s^{-1/2-\varepsilon})$ since $\int_s^t g_r dr = O_M(1)$, therefore the effect of this term can be merged into the final residual term of (4.24). Thus we get

$$x_t - \overline{y}_t = I_{1,t}^* + J_1 + J_2 + J_3 + O_M(s^{-1/2-\varepsilon}). \tag{4.25}$$

We show that $J_1 + J_2 + J_3 = O_M(s^{-1})$.
Estimation of J_1. To estimate J_1 write it as

$$J_1 = \int_s^t g_r y(r,s,x_s) dr = L_1(x_s), \quad \text{with} \quad L_1(x) = \int_s^t g_r y(r,s,x) dr.$$

Note that $L_1(0) = 0$. To estimate $L_1(x_s)$ consider a Taylor-series expansion of $L_1(x)$ around 0:

$$L_1(x_s) = \int_0^1 L_{1x}(\lambda x_s) d\lambda \cdot x_s \tag{4.26}$$

where $L_{1x} = (\partial/\partial x) L_1(x)$. It is easy to see that in computing L_{1x} differentiation and integration can be interchanged, thus we can write

$$L_{1x}(x) = \int_s^t \left(\frac{\partial}{\partial \xi} y(t,r,0) \cdot \frac{1}{r} \frac{\partial}{\partial x} \overline{H}(r,0,\omega) + \frac{\partial^2}{\partial \xi^2} y(t,r,0) \cdot \frac{1}{r} \overline{H}(r,0,\omega) \right) \cdot \frac{\partial}{\partial x} y(r,s,x)\ dr. \quad (4.27)$$

Since $y_\xi(t,r,0)$, $y_{\xi\xi}(t,r,0)$ and $y_x(r,s,x)$ are deterministic and bounded and $\overline{H}_x(r,0,\omega)$ and $\overline{H}(r,0,\omega)$ are zero-mean L-mixing processes we get by the moment inequality given as Theorem 8.1 that for each fixed $x \in D_0$

$$L_{1x}(x) = O_M\left(\int_s^t \frac{1}{r^2} dr \right)^{1/2} = O_M(s^{-1/2}).$$

Using similar arguments and taking into account that G is three-times continuously differentiable, we obtain that $L_{1xx}(x) = (\partial^2/\partial x^2) L_1(x) = O_M(s^{-1/2})$. Using now the maximal inequality given as Theorem 8.3 of the Appendix we get

$$\|L_{1x}(\lambda x_s)\| \leq \sup_{x \in D_0^*} \|L_{1x}(x)\| = O_M(s^{-1/2}).$$

Taking into account that $x_s = O_M(s^{-1/2})$ we finally get

$$J_1 = L_1(x_s) = O_M(s^{-1}). \quad (4.28)$$

Estimation of J_2. To estimate J_2 let us use the definition of $I_{1,t}^*$ and write

$$J_2 = \int_s^t g_r \int_s^r f_{1,v} \overline{H}(v,0,\omega) dv,$$

where the modulating function $f_{1,v}$ is

$$f_{1,v} = \frac{\partial}{\partial \xi} y(r,v,0) \cdot \frac{1}{v}.$$

Write g_r as

$$g_r = f_{2,1,r} \frac{\partial}{\partial x} \overline{H}(r,0,\omega) + f_{2,2,r} \overline{H}(r,0,\omega),$$

where the modulating functions are

$$f_{2,1,r} = \frac{\partial}{\partial \xi} y(t,r,0) \cdot \frac{1}{r} \ \text{ and } \ f_{2,2,r} = \frac{1}{r} \frac{\partial^2}{\partial \xi^2} y(t,r,0) \cdot \frac{1}{r}.$$

Noting that $y_\xi(t,r,0)$ and $y_{\xi\xi}(t,r,0)$ are bounded and applying the moment inequality for double integrals of L-mixing processes, given as Theorem 8.2 in the Appendix, we get

$$J_2 = O_M(s^{-1}). \quad (4.29)$$

Estimation of J_3. For J_3 we get after interchanging the order of integration:

$$J_3 = \int_s^t g_p x_p \left(\int_p^t g_r dr \right)\ dp.$$

For the inner integral we have

$$\int_p^t g_r dr = O_M(p^{-1/2})$$

by the moment inequality given as Theorem 8.1. Since $g_p = O_M(p^{-1})$ we have $g_p x_p = O_M(p^{-3/2})$ and thus the integrand of the outer integral is of the order of magnitude $O_M(p^{-2})$. It follows that

$$J_3 = O_M(s^{-1}). \quad (4.30)$$

Thus we conclude that indeed $J_1 + J_2 + J_3 = O_M(s^{-1})$ and substituting this into (4.25) the proof of Lemma 4.3 is complete.

Pasting together. Let us now take a subdivision of the half-line $[1,\infty)$ by the points q^i with $q > 1$ and let us consider an interval $q^n \le t < q^{n+1}$. Let us define for $i \ge 1$

$$\delta_i = I^*_{1,q^i,q^{i-1}} = \int_{q^{i-1}}^{q^i} \frac{\partial}{\partial \xi} y(q^i, r, 0) \frac{1}{r} \overline{H}(r, 0, \omega) dr.$$

Note that $\delta_i = O_M(q^{-i/2})$.

Lemma 4.4 *We have for $q^n \le t < q^{n+1}$*

$$x_t - y_t = \sum_{i=1}^{n} \frac{\partial}{\partial \xi} y(t, q^i, 0)\delta_i + I^*_{1,t,q^n} + O_M(q^{-n(1/2+\varepsilon_x)}). \tag{4.31}$$

Proof: Using Lemma 8.7 of the Appendix with $s_i = q^i, i = 0, 1, ..., n,\ s_{n+1} = t$ we get

$$x_t - y_t = \sum_{i=1}^{n} \int_0^1 \frac{\partial}{\partial \xi} y(t, q^i, w(\lambda))\ d\lambda \cdot \left(x_{q^i} - y(q^i, q^{i-1}, x_{q^{i-1}}) \right) + \left(x_t - y(t, q^n, x_{q^n}) \right), \tag{4.32}$$

where $w(i, \lambda) = (1 - \lambda) y(q^i, q^{i-1}, x_{q^{i-1}}) + \lambda x_{q^i}$. Taking into account Lemma 4.3 write the i-th local tracking error $x_{q^i} - y(q^i, q^{i-1}, x_{q^{i-1}})$ in the form $I^*_{1,q^i,q^{i-1}} + O_M(q^{-(i-1)(1/2+\varepsilon)}) = \delta_i + O_M(q^{-(i-1)(1/2+\varepsilon)})$ to get

$$x_t - y_t = \sum_{i=1}^{n} \int_0^1 \frac{\partial}{\partial \xi} y(t, q^i, w(\lambda)) d\lambda \cdot \left(\delta_i + O_M(q^{-(i-1)(1/2+\varepsilon)}) \right) + I^*_{1,t,q^n} + O_M(q^{-n(1/2+\varepsilon)}). \tag{4.33}$$

To estimate the cumulative effect of the error terms $O_M(q^{-(i-1)(1/2+\varepsilon)})$ note that we have $\|y_\xi(t, q^i, w(i, \lambda))\| \le C_0 (q^i/t)^\alpha \le C_0 (q^i/q^n)^\alpha = C_0 q^{-\alpha(n-i)}$, thus we get an upper bound

$$\sum_{i=1}^{n} \int_0^1 C_0 q^{-\alpha(n-i)} \cdot O_M(q^{-(i-1)(1/2+\varepsilon)}) + O_M(q^{-n(1/2+\varepsilon)}). \tag{4.34}$$

This expression can be estimated from above by using the remark after Lemma 8.5 of the Appendix, given as (8.6), applied for the sequences $(q^{-\alpha i})$ and $(q^{-(1/2+\varepsilon)i})$, the convolution of which is bounded from above by $C\ \max(q^{-\alpha n}, q^{-(1/2+\varepsilon)n})$ assuming that $\alpha \ne 1/2 + \varepsilon$. Since $\max(-\alpha, -(1/2 + \varepsilon)) = -\min(\alpha, 1/2 + \varepsilon) = -(1/2 + \min(\overline{\alpha}, \varepsilon))$ we get

$$x_t - y_t = \sum_{i=1}^{n} \int_0^1 \frac{\partial}{\partial \xi} y(t, q^i, w(i, \lambda)) d\lambda \cdot \delta_i + I^*_{1,t,q^n} + O_M(q^{-n(1/2+\varepsilon_x)}). \tag{4.35}$$

To further simplify the right hand side of (4.33) we replace $w(i, \lambda)$ by 0. Note that by Lemma 8.8 of the Appendix

$$\| \frac{\partial}{\partial \xi} y(t, q^i, w(i, \lambda)) - \frac{\partial}{\partial \xi} y(t, q^i, 0) \| \le C_0' (q^i/t)^\alpha |w(i, \lambda)|,$$

and hence cumulative error of this approximation is majorized by $C_0' \sum_{i=1}^{n} (q^i/t)^\alpha \cdot |w(i, \lambda)| \cdot \delta_i$. Note that $w(i, \lambda) = O_M(q^{-i/2})$, uniformly in λ since $x_{q^i} = O_M(q^{-i/2})$ by Theorem 3.1 and $|\overline{y}_{q^i}| = |y(q^i, q^{i-1}, x_{q^{i-1}})| \le C_0 |x_{q^{i-1}}|$, therefore $w(i, \lambda) \cdot \delta_i = O_M(q^{-i})$ uniformly in λ. Thus the the cumulative error of the last approximation is bounded from above by

$$C_0' \sum_{i=1}^{n} (q^i/t)^\alpha \cdot O_M(q^{-i}) \le \sum_{i=1}^{n} q^{-\alpha(n-i)} \cdot O_M(q^{-i}) = O_M(q^{-n\alpha'})$$

with $\alpha' = \min(\alpha, 1)_-$, by the remark after Lemma 8.5, given as (8.6). Since $1/2+\varepsilon < 1$ we have $\alpha' \geq \min(\alpha, 1/2+\varepsilon)_- = \varepsilon_x$ and with this the proof of the lemma is complete.

Now the i-th term on right hand side of (4.31) can be written as

$$\frac{\partial}{\partial\xi}y(t,q^i,0)\int_{q^{i-1}}^{q^i}\frac{\partial}{\partial\xi}y(q^i,r,0)\frac{1}{r}\overline{H}(r,0,\omega)dr = \int_{q^{i-1}}^{q^i}\frac{\partial}{\partial\xi}y(t,r,0)\frac{1}{r}\overline{H}(r,0,\omega)dr, \tag{4.36}$$

thus the cumulative contribution of the dominant terms in (4.31) is exactly what is the dominant term in Theorem 4.1. Since $y_t = O(t^{-\alpha}) = O(t^{-(1/2+\overline{\alpha})})$ the term can be merged into the residual term $O_M(q^{-n(1/2+\varepsilon_x)})$ and thus the proof of Theorem 4.1 has been completed.

Proof of Theorem 4.2: Let $(H^c(t,x,\omega))$ be the piecewise constant extension of $(H(n,x,\omega))$ defined under (3.35) and define a piecewise linear extension of (x_n) by

$$x_t^l = (t-n)x_n + (n+1-t)x_{n-1} \quad \text{for} \quad 1 \leq n \leq t \leq n+1 \quad \text{if} \quad x_{n-} \in \text{int}D_0. \tag{4.37}$$

On the other hand if $x_{n-} \notin \text{int}D_0$ then we reset x^l to its initial value ξ at time $t = n$ and put a hold on the recursion until $t = n+1$, i.e. we set

$$x_t^l = \xi \quad \text{for} \quad 1 \leq n < t \leq n+1 \quad \text{if} \quad x_{n-} \notin \text{int}D_0. \tag{4.38}$$

Note the shift in time: $x_1^l = x_0$.

Now it is easy to see that in intervals $n \leq t \leq n+1$ where no resetting takes place (x_t^l) satisfies a differential equation of the form

$$\dot{x}_t^l = \frac{1}{t}(H^c(t,x,\omega) + \delta H(t,\omega)), \tag{4.39}$$

where $\delta H(t,\omega) = \delta H^c(n,\omega) + O_M(t^{-1})$, cf. [19], (2.13). The conditions of Theorem 4.1 can be easily verified for the above procedure, except that we use the alternative resetting mechanism given by (3.17) and (3.19). Thus we get by Theorem 4.1

$$x_t^l - x^* = \int_1^t \frac{\partial}{\partial\xi}y(t,s,x^*)\frac{1}{s}H^c(s,x^*,\omega)ds + O_M(t^{-1/2-\varepsilon_x}). \tag{4.40}$$

Let $t = N$ be an integer and let $n \leq s < n+1$, with n being integer. We have

$$||\frac{\partial}{\partial\xi}y(t,s,x^*) - \frac{\partial}{\partial\xi}y(t,n,x^*)|| \leq C_0'(\frac{s}{t})^\alpha\frac{1}{t}. \tag{4.41}$$

Indeed, by Lemma 8.8

$$||y_{r\xi}(t,r,x^*)|| \leq C_0'(r/t)^\alpha \cdot ||\frac{1}{t}G_\xi(x^*)||.$$

Integrating $y_{r\xi}(t,r,\xi)$ between n and s we get (4.41). Now replacing $y_\xi(t,s,x^*)$ by $y_\xi(t,n,x^*)$ in (4.40), noting that $\frac{1}{s} - \frac{1}{n} = O(\frac{1}{s^2})$ and taking into account that $\overline{H}^c(t,x^*,\omega)$ is M-bounded we get that the cumulative error is of the order of magnitude

$$O_M\left(\int_1^t (\frac{s}{t})^\alpha \cdot \frac{1}{s^2}\ ds\right) = O_M(t^{-1}), \tag{4.42}$$

which can be merged into the residual term $O_M(t^{-1/2-\varepsilon_x})$ and thus the proof of Theorem 4.2 is complete.

Proof of Theorem 4.3: Defining H and δH as in (3.46) and (3.56) the conditions of Theorem 4.1 have been verified in Section 5 of [19]. In particular, the critical Condition 3.5 is verified in Lemma 5.6 in [19], (restated as Lemma 3.2 in the present paper), thus the claim follows.

5 The transformed error process is L-mixing

In this section we derive a useful corollary of Theorem 4.1, stating that an appropriate transformation of the error process $x_t - x^*$ is L-mixing. Define the transformed process

$$\tilde{x}_r = e^{r/2}(x_{e^r} - x^*). \tag{5.1}$$

The weak limit of the shifted process $(\tilde{x}_{r+\rho})$, when $\rho \to \infty$ is established in [5] and Theorem 13, Chapter 4.5, Part II of [3], under conditions, which are different from the conditions of the present paper. It is proven that $(\tilde{x}_{r+\rho})$ converges weakly to the solution of the linear stochastic differential equation

$$d\tilde{z}_r = (A^* + I/2)\tilde{z}_r + d\tilde{w}_r, \tag{5.2}$$

with zero initial condition, in short

$$(\tilde{x}_{r+\rho}) \to (\tilde{z}_r) \tag{5.3}$$

in weak sense, where, cf. (3.24),

$$A^* = \frac{\partial G(x)}{\partial x}\Big|_{x=x^*}$$

assuming that $(A^* + I/2)$ is stable. Here $d\tilde{w}_r$ is the stochastic differential of a Wiener-process, with some covariance matrix P^*dt. The weak limit $(\tilde{z}_r)$ is an L-mixing process with respect to the pair of σ-algebras $(\tilde{\mathcal{F}}_r, \tilde{\mathcal{F}}_r^+)$ generated by the past and future increments of the Wiener-process $(\tilde{w}_r)$, respectively. Hence it is indicated, but not implied by (5.3) that the transformed process $(\tilde{x}_r)$ itself is also L-mixing with respect to some pair of σ-algebras $(\tilde{\mathcal{F}}_r, \tilde{\mathcal{F}}_r^+)$. We prove that this is indeed the case.

The emphasis is on the non-asymptotic nature of our result. An analogous result for off-line prediction error estimators of ARMA-parameters has been proved in [24]. It extends to RPE estimators due to the strong approximation result given in result [22]. It has also been shown in [24] that this result is instrumental in deriving a pathwise characterization of performance degradation of an on-line adaptive predictor. Like in Section 4, we assume that $\delta G(t, y) = 0$, which implies $\mathrm{E}H(s, x^*, \omega) = 0$ exactly for all s.

Theorem 5.1 *Consider the continuous-time recursive estimation scheme given by (3.16) with the resetting mechanism (3.17) and (3.18). Assume that the conditions of Theorem 4.1 are satisfied. Then the transformed process $(\tilde{x}_r)$ is L-mixing with respect to $(\mathcal{F}_{e^r}, \mathcal{F}_{e^r}^+)$.*

Proof: *Approximation, dynamic representation and discretization of the process $(\tilde{x}_r)$.* By Theorem 4.1 the dominant term in the error process is

$$\int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H(s, x^*, \omega) ds,$$

and the error term is $O_M(t^{-1/2+\varepsilon_x})$. We transform the dominant term and the residual term $O_M(t^{-1/2+\varepsilon_x})$ in the same way as the error process itself (cf. (5.1)): we multiply by $t^{1/2}$ and introduce the new variables $t = e^r$ and $s = e^p$.

Now, since $\frac{\partial}{\partial \xi} y(t, s, x^*)$ is the solution of the variational equation

$$\frac{\partial}{\partial t} \frac{\partial}{\partial \xi} y(t, s, x^*) = \frac{1}{t} A^* \frac{\partial}{\partial \xi} y(t, s, x^*)$$

with initial condition $\frac{\partial}{\partial \xi} y(s, s, x^*) = I$, we get, using an exponential change of time-scale followed by an inverse change of time-scale, that

$$\frac{\partial}{\partial \xi} y(t, s, x^*) = e^{A^* \log(t/s)}.$$

Thus the dominant term in the error process gets transformed into

$$\tilde{x}_{1,r} = e^{r/2} \int_0^r e^{A^*(r-p)} H(e^p, x^*, \omega) dp. \tag{5.4}$$

Now Theorem 4.1 implies that the following:

Claim. *We have*

$$\tilde{x}_r - \tilde{x}_{1,r} = O_M(e^{-\varepsilon_x r}). \tag{5.5}$$

A dynamic representation of $\tilde{x}_{1,r}$ is obtained by differentiating (5.4) with respect to r. Then we get that $\tilde{x}_{1,r}$ satisfies the differential equation:

$$\frac{d}{dr}\tilde{x}_{1,r} = (A^* + I/2)\tilde{x}_{1,r} + e^{r/2}H(e^r, x^*, \omega), \qquad r \geq 0. \tag{5.6}$$

The dynamics satisfied by $(\tilde{x}_{1,r})$ is similar to the dynamics satisfied by (z_r), given by (5.2), but the process $e^{r/2}H(e^r, x^*, \omega)$ is not a good approximation to the increments of a Wiener-process, it is not even M-bounded. This difficulty can be avoided using discretization and averaging. Take a small, fixed positive number h and consider the discrete-time sampled process $\tilde{x}_{1,nh}$. It satisfies the discrete-time dynamics

$$\tilde{x}_{1,(n+1)h} = e^{(A^*+I/2)h}\tilde{x}_{1,nh} + \int_{nh}^{(n+1)h} e^{(A^*+I/2)((n+1)h-p)}e^{p/2}H(e^p, x^*, \omega)dp.$$

Note, that the input process is obtained as a weighted average of the input process of (5.6) over the interval $[nh, (n+1)h]$. Denote the second term on the right hand side, which is the input process for the discretized system, by $(\tilde{u}_{1,n})$, i.e. set

$$\tilde{u}_{1,n+1} = \int_{nh}^{(n+1)h} e^{(A^*+I/2)((n+1)h-p)}e^{p/2}H(e^p, x^*, \omega)dp. \tag{5.7}$$

The discrete-time dynamics:

$$\tilde{x}_{1,(n+1)h} = e^{(A^*+I/2)h}\tilde{x}_{1,nh} + \tilde{u}_{1,n+1}, \qquad n \geq 0, \tag{5.8}$$

with zero initial condition. In what follows we develop a series of approximations of the process $\tilde{u}_{1,n+1}$.

An averaging effect for the process $(\tilde{u}_{1,n})$. Going back to the original time-scale in (5.7) we can write $\tilde{u}_{1,n+1}$ as

$$\tilde{u}_{1,n+1} = \int_{e^{nh}}^{e^{(n+1)h}} \left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}}H(s, x^*, \omega)ds. \tag{5.9}$$

Claim U1. *For the order of magnitude of* $(\tilde{u}_{1,n+1})$ *we have*

$$\tilde{u}_{1,n+1} = O_M(h^{1/2}). \tag{5.10}$$

Indeed, using the moment inequality given as Theorem 8.1 we get that for any $q \geq 1$

$$\mathrm{E}^{1/q}|\tilde{u}_{1,n+1}|^q \leq C_q\left(\int_{e^{nh}}^{e^{(n+1)h}} ||\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}}||^2 ds\right)^{1/2} \cdot M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)).$$

Note that for $e^{nh} \leq s \leq e^{(n+1)h}$, $0 < h \leq h_0$ with some $0 < h_0$ fixed.

$$||\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}|| = ||e^{(p-(n+1)h)(-A^*-I/2)}|| \leq C, \tag{5.11}$$

where C is independent of n and h, since the set of matrices $e^{(p-(n+1)h)(-A^*-I/2)}$ with p varying between nh and $(n+1)h$ is compact. Thus we get

$$\mathrm{E}^{1/q}|\tilde{u}_{1,n+1}|^q \leq C_q\left(\int_{e^{nh}}^{e^{(n+1)h}} C^2\frac{1}{s}ds\right)^{1/2} \cdot M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)),$$

and the right hand side is $O(h^{1/2})$ indeed, as stated.

Truncated averaging: the process $(\tilde{u}_{2,n})$ and choosing δ_n and ε_δ. To eliminate the dependence in the process $(\tilde{u}_{1,n})$ we follow standard procedures, as described e.g. in [42]. First we remove a small portion of the integral by decreasing the upper limit of the integration to $e^{(n+1)h}-\delta_{n+1}$ with some positive δ_{n+1}. Since the original range of the integration has length $e^{(n+1)h}-e^{nh} = O(he^{nh})$ a reasonable choice for δ_{n+1} is

$$\delta_{n+1} = he^{\varepsilon_\delta nh}. \tag{5.12}$$

with $0 < \varepsilon_\delta < 1$. Thus we define

$$\tilde{u}_{2,n+1} = \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} H(s, x^*, \omega) ds. \tag{5.13}$$

Claim U2. *We have for $h > 0$*

$$\tilde{u}_{1,n+1} - \tilde{u}_{2,n+1} = O_M(h^{1/2}e^{-(1-\varepsilon_\delta)nh/2}) \quad \text{and} \quad \tilde{u}_{1,n+1} - \tilde{u}_{2,n+1} = O_M(h^{1/2}). \tag{5.14}$$

For the proof first note that

$$e^{(n+1)h} - \delta_{n+1} \geq e^{nh}. \tag{5.15}$$

Indeed, this is equivalent to $\delta_{n+1} \leq e^{(n+1)h} - e^{nh} = e^{nh}(e^h - 1)$ and since $\delta_{n+1} < he^{nh}$ and $h < (e^h - 1)$, the validity of (5.15) follows.

The error of the approximation is

$$\tilde{u}_{1,n+1} - \tilde{u}_{2,n+1} = \int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}} \left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} H(s, x^*, \omega) ds,$$

which can be estimated by the moment inequality given as Theorem 8.1. Thus we get that $\mathrm{E}^{1/q}|\tilde{u}_{1,n+1} - \tilde{u}_{2,n+1}|^q$ is bounded from above by

$$C_q \left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}} || \left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} ||^2 ds \right)^{1/2} \cdot M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*)).$$

Taking into account the kernel estimate given above as (5.11) we get that

$$\mathrm{E}^{1/q}|\tilde{u}_{1,n+1} - \tilde{u}_{2,n+1}|^q \leq C_q \left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}} \frac{C^2}{s} ds \right)^{1/2} \cdot M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*)). \tag{5.16}$$

For the integral term we have

$$\left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}} \frac{1}{s} ds \right)^{1/2} = (\log e^{(n+1)h} - \log(e^{(n+1)h} - \delta_{n+1}))^{1/2},$$

which is majorized by

$$\left(\delta_{n+1}/(e^{(n+1)h} - \delta_{n+1}) \right)^{1/2}.$$

Since $e^{(n+1)h} - \delta_{n+1} \geq e^{nh}$, we can continue the above inequality to get

$$\left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}} \frac{1}{s} ds \right)^{1/2} \leq \left(\delta_{n+1}/e^{nh} \right)^{1/2}.$$

Taking into account the definition of δ_{n+1} we get

$$\left(\delta_{n+1}/e^{nh} \right)^{1/2} = \left(he^{\varepsilon_\delta nh}/e^{nh} \right)^{1/2} = h^{1/2} e^{(\varepsilon_\delta - 1)nh/2}.$$

Combining the latter inequalities with (5.16) we get the first part of Claim U2, given as (5.14), while the second part is a trivial consequence.

The independent sequence $(\tilde{u}_{3,n})$. This is a key step in our arguments. We complete the construction of an approximating process of $(\tilde{u}_{1,n})$ by projecting $\tilde{u}_{2,n+1}$ on the relative future $\mathcal{F}^+_{e^{nh}-\delta_n}$. In fact, assuming that the conditional expectation operator and integration can be interchanged, we define

$$\tilde{u}_{3,n+1} = \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} \mathrm{E}(H(s,x^*,\omega)|\mathcal{F}^+_{e^{nh}-\delta_n})ds. \tag{5.17}$$

It is obvious that $(\tilde{u}_{3,n})$ constitutes an independent sequence of random variables adapted to $\mathcal{F}_{e^{nh}}$.

Remark: We will now approximate the process $\tilde{u}_{2,n+1}$ and get two dual error bounds. The first error bound ensures that the error is exponentially decaying, but there is multiplicative factor h^{-c} with $c > 0$, while the second bound ensures that the approximating process itself is of the order $O_M(h^{1/2})$.

Claim U3. *We have with $c > 0$ that shows up in Condition 3.1 (see the definition of L^+-mixing), the following two estimates:*

$$\tilde{u}_{2,n+1} - \tilde{u}_{3,n+1} = O_M(h^{-c}e^{-(1/2+c\varepsilon_\delta)nh}) \quad \text{and} \quad \tilde{u}_{2,n+1} - \tilde{u}_{3,n+1} = O_M(h^{1/2}). \tag{5.18}$$

First we show that $(\tilde{u}_{3,n+1})$ is an M-bounded sequence. Indeed, write $\tilde{u}_{3,n+1}$ as

$$\tilde{u}_{3,n+1} = \mathrm{E}\left(\int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} H(s,x^*,\omega)ds \;\middle|\; \mathcal{F}^+_{e^{nh}-\delta_n}\right)$$

and estimate the L_q-norm of the right hand side using Jensen's inequality. Taking into account (5.10), modified so that upper limit of the integration is reduced to the non-random upper limit $e^{(n+1)h} - \delta_{n+1}$, we get the claimed M-boundedness of $(\tilde{u}_{3,n+1})$ and in fact we get

$$\tilde{u}_{3,n+1} = O_M(h^{1/2}), \tag{5.19}$$

and thus the second part of the Claim U3 is proved.

To bound the approximation error more accurately define for $e^{nh} \le s \le e^{(n+1)h} - \delta_{n+1}$

$$v_s = |\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} \left(H(s,x^*,\omega) - \mathrm{E}(H(s,x^*,\omega)|\mathcal{F}^+_{e^{nh}-\delta_n})\right)|.$$

Then obviously

$$|\tilde{u}_{2,n+1} - \tilde{u}_{3,n+1}| \le \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} v_s ds.$$

and by the triangle inequality for the L_q-norm for $q \ge 1$

$$\mathrm{E}^{1/q}|\tilde{u}_{2,n+1} - \tilde{u}_{3,n+1}|^q \le \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \mathrm{E}^{1/q} v_s^q ds. \tag{5.20}$$

To estimate v_s note that $||(s/e^{(n+1)h})^{(-A^*-I/2)}|| \le C$ with some C for all s, n and h with $0 < h \le h_0$ and $s^{-1/2} \le e^{-nh/2}$. On the other hand we have for any $q \ge 1$

$$\mathrm{E}^{1/q}|(H(s,x^*,\omega) - \mathrm{E}(H(s,x^*,\omega)|\mathcal{F}^+_{e^{nh}-\delta_n}))|^q \le \gamma_q(s - (e^{nh} - \delta_n), H(x^*)),$$

and thus

$$\mathrm{E}^{1/q} v_s^q \le Ce^{-nh/2}\gamma_q(s - (e^{nh} - \delta_n), H(x^*)).$$

It follows that

$$\mathrm{E}^{1/q}|\tilde{u}_{3,n+1} - \tilde{u}_{2,n+1}|^q \le Ce^{-nh/2} \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \gamma_q(s - (e^{nh} - \delta_n), H(x^*))ds. \tag{5.21}$$

By Condition 3.1 $\gamma_q(s - (e^{nh} - \delta_n), H(x^*)) \le C(1 + \delta_n)^{-1-c}$. Furthermore note that the range of $\tau(s) = s - (e^{nh} - \delta_n)$ is included in the semi-infinite interval $[\delta_n, \infty)$, thus

$$\int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \gamma_q(s - (e^{nh} - \delta_n), H(x^*))ds \le \int_{\delta_n}^{\infty} \gamma_q(\tau, H(x^*))d\tau \le C'(1 + \delta_n)^{-c} < C'\delta_n^{-c}.$$

Combining this with (5.21) and taking into account the definitions of the lag δ_n given by (5.12) we get for any $1 \le q < \infty$

$$\mathrm{E}^{1/q}|\tilde{u}_{2,n+1} - \tilde{u}_{3,n+1}|^q \le C_q e^{-nh/2} h^{-c} e^{-c\varepsilon_\delta nh}$$

with some C_q, which is independent of n and h and this is equivalent to the first part of Claim U3, given as (5.18).

The final approximating process $(\tilde{x}_{3,nh})$. We are going to define a final approximation to $\tilde{x}_{nh}$ that plays a key role in subsequent analysis. This is obtained from the discrete-time dynamics (5.8) so that $\tilde{u}_{1,n+1}$ is replaced by $\tilde{u}_{3,n+1}$. Thus we define the process $(\tilde{x}_{3,(n+1)h})$ by

$$\tilde{x}_{3,(n+1)h} = e^{(A^* + I/2)h}\tilde{x}_{3,nh} + \tilde{u}_{3,n+1}, \qquad n \ge 0, \tag{5.22}$$

with zero initial condition. Let

$$\tilde{\mathcal{F}}_r = \mathcal{F}_{e^r} \quad \text{and} \quad \tilde{\mathcal{F}}_r^+ = \mathcal{F}_{e^r}^+.$$

We claim that the approximating process $(\tilde{x}_{3,nh})$ is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}_{nh}^+)$.

Indeed, since the real parts of the eigenvalues of A^* are less than or equal to α^* and $\alpha < \alpha^*$ the spectral norm of $e^{(A^*+I/2)h}$ is less than $e^{-\overline{\alpha}h}$ and hence there exists a $C > 0$ such that for any positive integer m

$$||e^{(A^*+I/2)mh}|| \le Ce^{-\overline{\alpha}mh}. \tag{5.23}$$

The input process, $(\tilde{u}_{3,n})$ is an M-bounded, independent, $\tilde{\mathcal{F}}_{nh}$-adapted sequence, hence it is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}_{nh}^+)$. Thus the output-process $(\tilde{x}_{3,nh})$ is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}_{nh}^+)$, by Lemma 8.4, as stated.

To get an accurate bound for the estimation error $\tilde{x}_{nh} - \tilde{x}_{3,nh}$ let us introduce the notations

$$\varepsilon_{x2} = \min(\overline{\alpha}, (1 - \varepsilon_\delta)/2) \tag{5.24}$$

$$\varepsilon_{x3} = \min(\overline{\alpha}, 1/2 + c\varepsilon_\delta). \tag{5.25}$$

Obviously $\varepsilon_{x2}, \varepsilon_{x3} > 0$. To formulate the next result note that if (ξ_t) and (η_t) are stochastic processes such that $\xi_t = O_M(c_t)$ and $\eta_t = O_M(d_t)$, where $c_t, d_t > 0$, then, trivially,

$$\xi_t + \eta_t = O_M(c_t + d_t). \tag{5.26}$$

Lemma 5.1 *The final approximation error* $\tilde{x}_{(n+1)h} - \tilde{x}_{3,(n+1)h}$ *is given by*

$$\tilde{x}_{(n+1)h} - \tilde{x}_{3,(n+1)h} = O_M(e^{-\varepsilon_x nh} + h^{1/2}e^{-\varepsilon_{x2}nh} + h^{-c}e^{-\varepsilon_{x3}nh}) = O_M(1). \tag{5.27}$$

Proof: The proof is almost trivial. It is easy to see, using the moment inequality given as Theorem 8.1 that both $(\tilde{x}_{nh})$ and $(\tilde{x}_{3,nh})$ are M-bounded, which implies the second part of the claim. To prove the first part, first note that the first term on the right hand side comes from (5.5). Next note that the error process $(\tilde{x}_{1,(n+1)h} - \tilde{x}_{3,(n+1)h})$ satisfies

$$(\tilde{x}_{1,(n+1)h} - \tilde{x}_{3,(n+1)h}) = e^{(A^*+I/2)h}(\tilde{x}_{1,nh} - \tilde{x}_{3,nh}) + (\tilde{u}_{1,n+1} - \tilde{u}_{3,n+1}), \qquad n \ge 0$$

with zero initial conditions. For the input process $(\tilde{u}_{1,n+1} - \tilde{u}_{3,n+1})$ the combination of the upper bounds given in Claim U2 and 3 or equivalently in (5.14) and (5.18) is used. Applying Lemma 8.5 we get the second and third terms on the right hand side of (5.27), which is thus proved.

To complete the proof of Theorem 5.1 we first note that defining

$$r_n = \tilde{x}_{3,nh} - \tilde{x}_{nh},$$

this residual process is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}^+_{nh})$. Indeed, (r_n) is M-bounded and $\tilde{\mathcal{F}}_{nh}$-measurable. On the other hand, writing (5.27) in the form $r_n = O_M(e^{-\varepsilon'_x nh})$ we get for any integer $\tau \geq 0$

$$\gamma_q(\tau, r) = \sup_{n \geq \tau} \mathrm{E}^{1/q} |r_n - \mathrm{E}\ [r_n | \mathcal{F}^+_{n-\tau}]\ |^q \leq 2 \sup_{n \geq \tau} \mathrm{E}^{1/q} |r_n|^q \leq 2C_q e^{-\varepsilon'_x \tau h}, \tag{5.28}$$

with some finite C_q. The right hand side is obviously summable over τ and thus we get the claim.

Since the class of L-mixing processes is closed under addition, it follows that $\tilde{x}_{nh}$ is also L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}^+_{nh})$. The second remark we need is that the processes $(\tilde{x}_{nh+d})$ and $(\tilde{x}_{3,nh+d})$, with $0 \leq d < h$ fixed can be analyzed similarly and it is easy to see that all the relevant estimates are valid uniformly in d. Thus we conclude, that the processes $(\tilde{x}_{nh+d})$ are L-mixing with respect to $(\tilde{\mathcal{F}}_{nh+d}, \tilde{\mathcal{F}}^+_{nh+d})$, uniformly in d for $0 \leq d < h$. Applying Corollary 3.5 of [24], restated as Lemma 8.3 in the Appendix, implies that the continuous-time process $(\tilde{x}_r)$ itself is L-mixing with respect to $(\tilde{\mathcal{F}}_r, \tilde{\mathcal{F}}^+_r) = (\mathcal{F}_{e^r}, \mathcal{F}^+_{e^r})$ and the proof is complete.

6 The asymptotic covariance matrix

The asymptotic covariance matrix for Algorithm DFL, (3.53)-(3.54), has been rigorously derived in Theorem 13, Chapter 4.5, Part II of [3] in a *series* model, where the initial time tends to infinity, and thus the probability of exiting the truncation domain tends to 0. The asymptotic covariance-matrix of Robbins-Monroe type recursive estimators has been known for long time, cf. e.g. [54]. Here the correction term $H(n, x, \omega)$ is assumed to form an independent sequence, see Condition A.3 in Chapter 2.3 of [54]. The asymptotic covariance-matrix for the RPE estimator of ARMA-processes has been first given in [60] using the eventually false a priori assumption that the non-truncated estimator sequence converges almost surely. It is likely that the analysis of the cited paper carries over to truncated estimators.

The purpose of this section is to derive the asymptotic covariance matrix for the general continuous-time recursive estimator process Algorithm CR given in (3.16) equipped with a re-setting mechanism defined under (3.17) and (3.18). The study of the discrete time procedure Algorithm DR given in (3.34) and Algorithm DFL, given under (3.53)-(3.54), with resetting mechanisms defined in Section 3, can be reduced to the study of Algorithm CR, as pointed out in Section 3 and 4. The main advance of this section relative to the cited result of [3] is that the asymptotic covariance matrix for the DFL scheme with enforced boundedness is obtained for a *single* process.

We also get a rate of convergence for the covariance-matrix sequence, which is useful in applications such as the analysis of performance degradation to statistical parametric uncertainty. For the present section we need the following additional condition:

Condition 6.1 *We assume that $(H(s, x^*, \omega))$ is asymptotically wide-sense stationary in the following sense: there exists a zero-mean, wide-sense stationary process $(H_0(s, x^*, \omega))$ such that*

$$\eta_s = H(s, x^*, \omega) - H_0(s, x^*, \omega) = O_M(s^{-1-\varepsilon_H}) \tag{6.1}$$

with some $\varepsilon_H > 0$.

This condition is easily verified in system-identification. In fact, if we consider the general estimation scheme of Section 3 defined by (3.53)-(3.54), then it is easy to see that we have $\eta_s = O_M(e^{-\beta s})$ with some $\beta > 0$. Now we have the following modification of Theorem 4.1:

Theorem 6.1 *Consider the continuous-time recursive estimation scheme given by (3.16) with the resetting mechanism (3.17) and (3.18). Assume that the conditions of Theorem 4.1 are satisfied and in addition Condition 6.1 is also satisfied. Recall that* $\varepsilon_x = \min(\overline{\alpha}, \varepsilon)_-$, *where* $\overline{\alpha}$ *is defined under (3.31) and* ε *is given in Condition 3.5. Then we have*

$$x_t - x^* = \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H_0(s, x^*, \omega) ds + O_M(t^{-1/2-\varepsilon_x}), \tag{6.2}$$

and the wide-sense stationary process $(H_0(s, x^*, \omega))$ *is* L^+*-mixing.*

Proof: Consider the expression for the error $x_t - x^*$ that has been given in Theorem 4.1, or in (4.1). The difference between (4.1) and (6.2) is in the dominant terms and this difference can be majorized by

$$\int_1^t |\frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} \eta_s| ds \le \int_1^t C_0 (s/t)^\alpha |\frac{1}{s} \eta_s| ds,$$

due to Condition 3.4. Taking the L_q-norm of both sides with some $q \ge 1$ and applying the triangle inequality for L_q-norms we get an upper bound of the form

$$C_q \int_1^t C_0 (s/t)^\alpha \frac{1}{s} s^{-1-\varepsilon_H} ds,$$

which is majorized by $C'_q t^{-1-\varepsilon_H}$. Thus the difference between the dominant terms is certainly $O_M(t^{-1/2-\varepsilon_x})$ and thus (6.2) follows.

To prove that $(H_0(s, x^*, \omega))$ is L^+-mixing note that repeating the argument leading to (5.28) gives that for any integer $\tau \ge 0$

$$\gamma_q(\tau, \eta) \le 2 C_q \tau^{-1-\varepsilon_H},$$

and hence (η_s) is L^+-mixing. Since the class of L^+-mixing processes is closed under addition, it follows that $(H_0(s, x^*, \omega))$ is also L^+-mixing and the proof is complete.

Remark. There is no loss of generality to assume that

$$\gamma_q(\tau, H_0) \le C_q (1 + \tau)^{-1-c_q}$$

for all $\tau \ge 0$ with the same C_q, c_q as in Condition 3.1 requiring that H and $\Delta H / \Delta x$ be L^+ mixing.

To formulate the basic result of this section we need some notations. Denoting the autocovariance matrix of $H_0(s, x^*, \omega)$ by $\rho(\tau)$, i.e. setting

$$\rho(\tau) = \mathrm{E}\,[H_0(s + \tau, x^*, \omega) H_0^T(s, x^*, \omega)] = \mathrm{E}\,[H_0(\tau, x^*, \omega) H_0^T(0, x^*, \omega)],$$

we define a basic quantity:

$$P^* = \int_{-\infty}^{\infty} \rho(\tau) d\tau. \tag{6.3}$$

Since the process $(H_0(s, x^*, \omega))$ is L-mixing, the integral above converges. Indeed, since $H_0 = (H_0(s, x^*, \omega))$ is a wide-sense stationary zero-mean L-mixing process, using Lemma 8.1 with $p = q = 2$, we get

$$\rho(\tau) \le C \gamma_2(|\tau|, H_0) \tag{6.4}$$

with some $C > 0$, thus integrability follows.

It is easy to see, cf. Lemma 6.4 below, that the matrix P^* is the asymptotic covariance matrix of the arithmetic mean

$$\frac{1}{2T} \int_{-T}^{T} H_0(s, x^*, \omega) ds,$$

i.e. we have

$$P^* = \lim_{T \to \infty} 2T\ \mathrm{E}\,[\left(\frac{1}{2T} \int_{-T}^{T} H_0(s, x^*, \omega) ds\right) \left(\frac{1}{2T} \int_{-T}^{T} H_0(s, x^*, \omega) ds\right)^T]. \tag{6.5}$$

We will also need the notation introduced in (3.24):

$$A^* = \frac{\partial G(x)}{\partial x}|_{x=x^*}.$$

The value of the asymptotic covariance matrix can be easily guessed. Namely, the assumed validity of (5.3) implies that, $t^{1/2}(x_t - x^*)$ is asymptotically normally distributed with zero mean and covariance matrix S^*, which satisfies the Lyapunov-equation (6.6) below. This result on the asymptotic covariance-matrix of the estimator, has strong roots in the classical theory of stochastic approximation, see [54]. The closest to our result is Theorem 13, Chapter 4.5, Part II. of [3].

Theorem 6.2 *Consider the continuous-time recursive estimation scheme given by (3.16) with the resetting mechanism (3.17) and (3.18). Assume that the conditions of Theorem 4.1 are satisfied and in addition $(H(s, x^*, \omega))$ satisfies Condition 6.1. Then the asymptotic covariance-matrix of the error process $(x_t - x^*)$, defined by*

$$S^* = \lim_{t\to\infty} t\mathrm{E}[(x_t - x^*)(x_t - x^*)^T],$$

exists and it satisfies the Lyapunov-equation

$$(A^* + I/2)S^* + S^*(A^* + I/2)^T + P^* = 0, \tag{6.6}$$

where A^* is defined above, (see also (3.24)) and P^* is defined by (6.3) above. More exactly we have with some $\varepsilon_{xx} > 0$

$$\mathrm{E}[(x_t - x^*)(x_t - x^*)^T] = \frac{1}{t}S^* + O(t^{-1-\varepsilon_{xx}}).$$

Remark. In the case of a stochastic Newton method, i.e. when $A^* = -I$, we get

$$S^* = P^*.$$

In the context of Algorithm DFL, (3.53)-(3.54) this can be directly seen from Theorem 4.4.

Take the example of the recursive LSQ estimation of an AR(p) process given

$$y_n = (\theta^*)^T\phi_n + e_n,$$

where θ^* is the p-dimensional AR-parameter, $\phi_n = (-y_{n-1}, ..., -y_{n-p})^T$ and e_n is the noise term with variance $\sigma^2(e)$. AR-processes are special in the sense that the off-line LSQ estimator can be computed *exactly* in a recursive fashion, thus the off-line and on-line estimators, if properly initialized, coincide and their asymptotic covariance is the same. A non-trivial corollary of Theorem 5.2 is that this is still the case if both estimators are forced to stay inside a compact domain using truncation for the off-line estimator and resetting for the on-line estimator.

Let

$$R^* = \mathrm{E}\phi_n\phi_n^T$$

assuming stationarity of ϕ_n. Then, under well-known conditions the asymptotic covariance matrix of the LSQ-estimator is known to be

$$S^* = \sigma^2(e)(R^*)^{-1}.$$

For the RLSQ estimator we have the updating term, with $x = \theta$,

$$H_n(s, \theta, \omega) = R^{-1}\phi_n(y_n - \phi_n^T\theta)$$

from which we get

$$G(\theta) = \theta^* - \theta$$

thus the RLSQ-method is a stochastic Newton method. Since

$$H_n(s, \theta^*, \omega) = R^{-1}\phi_n e_n$$

we get

$$P^* = \sigma^2(e)(R^*)^{-1}$$

which indeed agrees with S^*.

Remark. For the discrete time method, Algorithm DR, we have

$$x_t^l - x^* = \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H^c(s, x^*, \omega) ds + O_M(t^{-1/2-\varepsilon_x}).$$

see (4.40) and the analysis given below is applicable. Note however that now we get the familiar expression, see [54],

$$P^* = \int_{-\infty}^{\infty} \mathrm{E}\ [H_0^c(\tau, x^*, \omega) H_0^{cT}(0, x^*, \omega)] d\tau = \sum_{-\infty}^{\infty} \mathrm{E}\ [H_0(m, x^*, \omega) H_0^T(0, x^*, \omega)]. \tag{6.7}$$

Proof: *Reduction to the process* $(\tilde{x}_{3,nh})$. The claim of the theorem can be reformulated in terms of the transformed process, with $t = e^r$, as follows: we have with some $\varepsilon_{xx} > 0$

$$\mathrm{E}[\tilde{x}_r \tilde{x}_r^T] = S^* + O(e^{-\varepsilon_{xx} r}). \tag{6.8}$$

Now by Lemma 5.1 we have with $r = nh$

$$x_t - x^* = \frac{1}{e^{r/2}} \tilde{x}_r = \frac{1}{e^{nh/2}} \tilde{x}_{3,nh} + \frac{1}{e^{nh/2}} O_M(e^{-\varepsilon_x nh} + h^{1/2} e^{-\varepsilon_{x2} nh} + h^{-c} e^{-\varepsilon_{x3} nh}). \tag{6.9}$$

Multiplying both sides $e^{nh/2}$, squaring them and taking into account the second part of Lemma 5.1, we get the following key lemma:

Lemma 6.1 *We have with* $r = nh$*:*

$$\mathrm{E}[\tilde{x}_r \tilde{x}_r^T] = \mathrm{E}[\tilde{x}_{3,nh} \tilde{x}_{3,nh}^T] + O(e^{-\varepsilon_x nh} + h^{1/2} e^{-\varepsilon_{x2} nh} + h^{-c} e^{-\varepsilon_{x3} nh}). \tag{6.10}$$

This error-estimate seems to be fragile, due to Terms 3 and 4 on the right hand side, in view of the fact that the left hand side is $O(h)$, but this weakness will be eliminated at the very end of the proof of Theorem 6.2 by appropriate choice of h.
Thus the study of the covariance-matrix of x_t is reduced to the study of the covariance matrix of $\tilde{x}_{3,nh}$, which will be denoted by $R_{3,n}^{\tilde{x}}$:

$$R_{3,n}^{\tilde{x}} = \mathrm{E}\ [\tilde{x}_{3,nh} \tilde{x}_{3,nh}^T].$$

Now change n for $n+1$ and note that $\tilde{x}_{3,(n+1)h}$ is defined via the discrete-time dynamical system (5.22), in which the input process $(\tilde{u}_{3,n+1})$ consists of a sequence of independent random variables. The covariance-matrix of $\tilde{u}_{3,m}$ will be denoted by

$$R_{3,m}^{\tilde{u}} = \mathrm{E}[\tilde{u}_{3,m} \tilde{u}_{3,m}^T].$$

In what follows we shall develop a sequence of approximations of $\tilde{u}_{3,m}$ to get a nice approximation for $R_{3,m}^{\tilde{u}}$.

The approximating process $(\tilde{u}_{4,m})$. Let us recall, see (5.9), that in the original time-scale we have

$$\tilde{u}_{1,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \left(\frac{s}{e^{(m+1)h}} \right)^{(-A^* - I/2)} \frac{1}{s^{1/2}} H(s, x^*, \omega) ds.$$

Approximate $\tilde{u}_{1,m+1}$ by replacing the kernel within the integrand by 1, i.e. set

$$\tilde{u}_{4,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \frac{1}{s^{1/2}} H(s, x^*, \omega) ds.$$

Claim U4. *We have*

$$\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1} = O_M(h^{3/2}). \tag{6.11}$$

To prove the claim note that the approximation error can be written as

$$\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \left(\left(\frac{s}{e^{(m+1)h}} \right)^{(-A^*-I/2)} - I \right) \frac{1}{s^{1/2}} H(s, x^*, \omega) ds.$$

Using the moment inequality given as Theorem 8.1 we get for any $q \geq 2$ that $\mathrm{E}^{1/q}|\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1}|^q$ is bounded from above by

$$C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} || \left(\left(\frac{s}{e^{(m+1)h}} \right)^{(-A^*-I/2)} - I \right) \frac{1}{s^{1/2}} ||^2 ds \right)^{1/2} M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*)).$$

Now, $\|(s/e^{(m+1)h})^{(-A^*-I/2)} - I\| = \|e^{(-A^*-I/2)(\log s - (m+1)h)} - I\| \leq ch$, with some c, which depends only on A^*, for $0 < h \leq h_0$, since $-h \leq (\log s - (m+1)h) \leq 0$. (Apply a Taylor-series expansion of the matrix-exponential to get the desired inequality). Thus we get

$$\mathrm{E}^{1/q}|\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1}|^q \leq C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} (ch)^2 \frac{1}{s} ds \right)^{1/2} M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*)).$$

and from here

$$\mathrm{E}^{1/q}|\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1}|^q \leq Ch^{3/2},$$

where C is independent of h and thus Claim U4 follows.

The approximating process $(\tilde{u}_{5,m+1})$. This approximation is obtained by replacing $\frac{1}{s^{1/2}}$ within the integral by a constant:

$$\tilde{u}_{5,m+1} = \frac{1}{e^{mh/2}} \int_{e^{mh}}^{e^{(m+1)h}} H(s, x^*, \omega) ds.$$

Claim U5. *We have*

$$\tilde{u}_{4,m+1} - \tilde{u}_{5,m+1} = O_M(h^{3/2}). \tag{6.12}$$

Indeed, we have

$$\tilde{u}_{5,m+1} - \tilde{u}_{4,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \left(\frac{1}{e^{mh/2}} - \frac{1}{s^{1/2}} \right) H(s, x^*, \omega) ds,$$

and we can apply the moment inequality Theorem 8.1. For this purpose we estimate the integrand:

$$0 \leq \left(\frac{1}{e^{mh/2}} - \frac{1}{s^{1/2}} \right) \leq \left(\frac{1}{e^{mh/2}} - \frac{1}{e^{(m+1)h/2}} \right) \leq \frac{1}{(e^{mh/2})^2} (e^{(m+1)h/2} - e^{mh/2})$$

$$\leq \frac{1}{e^{mh/2}} (e^{h/2} - 1) \leq \frac{1}{e^{mh/2}} h$$

for small h. Thus we get for $q \geq 2$

$$\mathrm{E}^{1/q}|\tilde{u}_{5,m+1} - \tilde{u}_{4,m+1}|^q \leq C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} \left(\frac{1}{e^{mh/2}} - \frac{1}{s^{1/2}} \right)^2 ds \right)^{1/2} M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*)) \leq$$

$$\leq \quad C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} \frac{h^2}{e^{mh}} ds \right)^{1/2} M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*)) = O(h^{3/2}),$$

and the claim follows.

The approximating process $(\tilde{u}_{6,m+1})$. This approximation is obtained by replacing $H(s,x^*,\omega)$ by $H_0(s,x^*,\omega)$ in the definition of $\tilde{u}_{6,m+1}$, i.e. we define

$$\tilde{u}_{6,m+1} = \frac{1}{e^{mh/2}} \int_{e^{mh}}^{e^{(m+1)h}} H_0(s,x^*,\omega)ds. \tag{6.13}$$

Claim U6. *We have*

$$\tilde{u}_{5,m+1} - \tilde{u}_{6,m+1} = O_M(he^{-mh(1/2+\varepsilon_H)}) \quad \text{and} \quad \tilde{u}_{5,m+1} - \tilde{u}_{6,m+1} = O_M(h^{1/2}). \tag{6.14}$$

Indeed, we have, using (6.1),

$$\tilde{u}_{5,m+1} - \tilde{u}_{6,m+1} = \frac{1}{e^{mh/2}} \int_{e^{mh}}^{e^{(m+1)h}} \eta_s ds =$$
$$= \frac{1}{e^{mh/2}}(e^{(m+1)h} - e^{mh})O_M(e^{-mh(1+\varepsilon_H)}) = O_M(he^{-mh(1/2+\varepsilon_H)})$$

and the first part of the claim follows. The second part is a direct consequence of Theorem 8.1.

Summarizing the equations expressing the approximation errors between the successive values $\tilde{u}_3, \tilde{u}_2, \tilde{u}_1, \tilde{u}_4, \tilde{u}_5, \tilde{u}_6$ given by (5.14), (5.18), (6.11), (6.12), (6.14) we get

Lemma 6.2 *Let c be as in Condition 3.1, requiring that H be L^+ mixing. Then we have*

$$\tilde{u}_{3,m} = \tilde{u}_{6,m} + O_M(h^{-c}e^{-(1/2+c\varepsilon_\delta)mh} + h^{1/2}e^{-(1-\varepsilon_\delta)mh/2} + +h^{3/2} + he^{-mh(1/2+\varepsilon_H)}), \tag{6.15}$$

and all error terms are also $O_M(h^{1/2})$.

Squaring this equation we get for $R^{\tilde{u}}_{3,m+1} = \mathrm{E}\,[\tilde{u}_{3,m}\tilde{u}^T_{3,m}]$

$$R^{\tilde{u}}_{3,m+1} = \mathrm{E}\,[\tilde{u}_{6,m}\tilde{u}^T_{6,m}] + O(h^{1/2-c}e^{-(1/2+c\varepsilon_\delta)mh} + O(he^{-(1-\varepsilon_\delta)mh/2} + h^2 + h^{3/2}e^{-mh(1/2+\varepsilon_H)}). \tag{6.16}$$

The covariance matrix of $\tilde{u}_{6,m+1}$. Next we show that the covariance matrix of the approximation $\tilde{u}_{6,m+1}$ can be expressed in terms of the matrix P^*. This is no surprise in view of the assumed validity of (6.5), but to capture the rate of convergence extra work is needed.

Lemma 6.3 *Let c be as in Condition 3.1, requiring that H be L^+ mixing. Then we have*

$$\mathrm{E}[\tilde{u}_{6,m+1}\tilde{u}^T_{6,m+1}] = hP^* + O(h^{1-c}e^{-cmh}). \tag{6.17}$$

Proof: Consider normalized arithmetic means of the form:

$$s_{A,B} = \frac{1}{(B-A)^{1/2}} \int_A^B H_0(s,x^*,\omega)ds$$

with $A < B$. It is obvious that

$$\mathrm{E}\,[s_{A,B}\ s^T_{A,B}] = \frac{1}{B-A} \int_A^B \int_A^B \rho(s-s')\ ds\ ds',$$

where $\rho(\tau)$ is the autocovariance function of the process $H_0 = (H_0(s,x^*,\omega))$.

Note that if $H_0 = (H_0(s,x^*,\omega))$ is a wide-sense stationary zero-mean L^+-mixing process then we have, using (6.4) and the inequality $\gamma_2(|\tau|, H_0) \le C(1+|\tau|)^{-c}$,

$$\rho(\tau) \le C(1+|\tau|)^{-c}. \tag{6.18}$$

with some $C, c > 0$. Applying Lemma 6.4 below with

$$A = e^{mh} \quad \text{and} B = e^{(m+1)h}$$

we have $(B-A) = e^{mh}(h+O(h^2))$ for small h. Thus we get, using the inequality $(1+B-A)^{-c} < C'(B-A)^{-c}$,

$$\frac{1}{B-A}\int_A^B \int_A^B \rho(s-s')\ ds\ ds' = P^* + O(e^{-cmh}h^{-c}),$$

and from here we get for the covariance-matrix of $\tilde{u}_{6,m+1}$

$$\mathrm{E}[\tilde{u}_{6,m+1}\tilde{u}_{6,m+1}^T] = \frac{1}{e^{mh}}\int_{e^{mh}}^{e^{(m+1)h}} \int_{e^{mh}}^{e^{(m+1)h}} \rho(s-s')dsds' = \frac{B-A}{e^{mh}}(P^* + O(e^{-cmh}h^{-c})),$$

and Lemma 6.3 follows.

Lemma 6.4 *Let $(\rho(\tau)), -\infty < \tau < \infty$ be a matrix-valued measurable function process, satisfying $||\rho(\tau)|| \leq C(1+|\tau|)^{-c}$ with some $C, c > 0$. Then we have for any $A < B$*

$$\frac{1}{B-A}\int_A^B \int_A^B \rho(s-s')ds\ ds' = P^* + O((1+B-A)^{-c}).$$

Proof: Introduce the new variables $\tau = s - s', \mu = s + s'$. This change of coordinates has a Jacobian with determinant 2, i.e. $d\tau\ d\mu = 2ds\ ds'$. The new variable τ takes it values between $-(B-A)$ and $B-A$ and for each fixed τ the possible values of μ are in the interval $(2A+|\tau|, 2B-|\tau|)$. Thus

$$\begin{aligned} P_{A,B} &= \frac{1}{B-A}\int_{-(B-A)}^{B-A}\int_{2A+|\tau|}^{2B-|\tau|} \rho(\tau)\frac{1}{2}d\tau\ d\mu = \\ &= \frac{1}{B-A}\int_{-(B-A)}^{B-A}(2B-2A-2|\tau|)\rho(\tau)\frac{1}{2}d\tau\ d\mu = \int_{-(B-A)}^{B-A}\left(1-\frac{|\tau|}{B-A}\right)\rho(\tau)d\tau. \end{aligned}$$

From here it follows immediately that

$$||P_{A,B}|| = ||\frac{1}{B-A}\int_A^B\int_A^B \rho(s-s')ds\ ds'|| \leq \int_{-(B-A)}^{B-A}||\rho(\tau)||d\tau \leq \int_{-\infty}^{\infty}||\rho(\tau)||d\tau. \qquad (6.19)$$

This inequality will be used subsequently. Now, write

$$\int_{-(B-A)}^{B-A}\left(1-\frac{|\tau|}{B-A}\right)\rho(\tau)d\tau = \int_{-(B-A)}^{B-A}\rho(\tau)d\tau - \int_{-(B-A)}^{B-A}\frac{|\tau|}{B-A}\rho(\tau)d\tau.$$

Then

$$\begin{aligned} \int_{-(B-A)}^{B-A}\rho(\tau)d\tau - P^* &= \int_{-(B-A)}^{B-A}\rho(\tau)d\tau - \int_{-\infty}^{\infty}\rho(\tau)d\tau = \\ &- \int_{-\infty}^{-(B-A)}\rho(\tau)d\tau - \int_{B-A}^{\infty}\rho(\tau)d\tau. \end{aligned}$$

Taking into account that $||\rho(\tau)|| \leq C(1+|\tau|)^{-1-c}$, we get that

$$||-\int_{-\infty}^{-(B-A)}\rho(\tau)d\tau - \int_{B-A}^{\infty}\rho(\tau)d\tau|| \leq \frac{2C}{c}(1+B-A)^{-c}. \qquad (6.20)$$

On the other hand

$$\int_{-(B-A)}^{B-A}\frac{|\tau|}{B-A}||\rho(\tau)||d\tau \leq \int_{-(B-A)}^{B-A}\frac{|\tau|}{B-A}C(1+|\tau|)^{-1-c}d\tau.$$

Write $|\tau|C(1+|\tau|)^{-1-c} \leq (1+|\tau|)C(1+|\tau|)^{-1-c} = C(1+|\tau|)^{-c}$ and use the symmetry of the last integrand above to get the upper bound

$$2\int_0^{B-A} \frac{1}{B-A} C(1+\tau)^{-c} d\tau \leq \frac{2}{B-A}\frac{C}{(-c+1)}(1+\tau)^{-c+1}|_0^{B-A} =$$
$$\frac{2}{B-A}\frac{C}{(-c+1)}((1+B-A)^{-c+1}-1)) \leq C'(1+B-A)^{-c}$$

and combining this with (6.20) the proposition of the lemma follows.

The final approximation of $R^{\tilde{u}}_{3,m+1}$: Combining (6.16) and (6.17) we get

$$R^{\tilde{u}}_{3,m+1} = \mathrm{E}\,[\tilde{u}_{3,m}\tilde{u}^T_{3,m}] = hP^* +$$
$$O(h^{1-c}e^{-cmh} + h^{1/2-c}e^{-(1/2+c\varepsilon_\delta)mh} + he^{-(1-\varepsilon_\delta)mh/2} + h^2 + h^{3/2}e^{-mh(1/2+\varepsilon_H)}). \quad (6.21)$$

To simplify notations write the residual terms in the form $h^{\beta_i}e^{-\varepsilon_i mh}, i = 1, ..., 5$, with

$$\begin{array}{ll} \beta_1 = 1-c, & \varepsilon_1 = c \\ \beta_2 = 1/2-c, & \varepsilon_2 = 1/2 + c\varepsilon_\delta \\ \beta_3 = 1, & \varepsilon_3 = (1-\varepsilon_\delta)/2 \\ \beta_4 = 2, & \varepsilon_4 = 0 \\ \beta_5 = 3/2, & \varepsilon_5 = 1/2 + \varepsilon_H. \end{array} \quad (6.22)$$

Obviously we have $\varepsilon_i > 0$ for $i \neq 4$. For $i = 4$ we have $\varepsilon_4 = 0$, but then $\beta_4 = 2$. With this notations we can formulate the following lemma:

Lemma 6.5 *We have with* $h^{\beta_i}e^{-\varepsilon_i mh}, i = 1, ..., 5$ *defined under (6.22)*

$$R^{\tilde{u}}_{3,m+1} = \mathrm{E}\,[\tilde{u}_{3,m}\tilde{u}^T_{3,m}] = hP^* + \sum_{i=1}^{5} O(h^{\beta_i}e^{-\varepsilon_i mh}). \quad (6.23)$$

The discrete-time Lyapunov-equation. Consider the discrete-time dynamics followed by $(\tilde{x}_{3,nh})$, given by (5.22). Since the input process is a sequence of independent random variables it follows that the covariance matrix of $\tilde{x}_{3,nh}$, denoted by $R^{\tilde{x}}_{3,n}$ $R^{\tilde{x}}_{3,n+1}$, satisfies the Lyapunov-equation:

$$R^{\tilde{x}}_{3,n+1} = e^{(A^*+I/2)h}R^{\tilde{x}}_{3,n}e^{(A^*+I/2)^Th} + R^{\tilde{u}}_{3,n+1},$$

with zero initial condition. Substituting $R^{\tilde{u}}_{3,n+1}$ from (6.23) and setting $n = m$, we get

$$R^{\tilde{x}}_{3,m+1} = e^{(A^*+I/2)h}R^{\tilde{x}}_{3,m}e^{(A^*+I/2)^Th} + hP^* + \sum_{i=1}^{5} O(h^{\beta_i}e^{-\varepsilon_i mh}). \quad (6.24)$$

Solving this iteratively in the range $0 \leq m \leq n$ we get

$$\begin{aligned} R^{\tilde{x}}_{3,n+1} &= \sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h}\ hP^*\ e^{(A^*+I/2)^T(n-m)h} + \\ &+ \sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h}\left(\sum_{i=1}^{5} h^{\beta_i}e^{-\varepsilon_i mh}\right)e^{(A^*+I/2)^T(n-m)h}. \end{aligned} \quad (6.25)$$

The contributions of the terms $h^{\beta_i}e^{-\varepsilon_i mh}, i = 1, ..., 5$ are estimated as follows:

$$\begin{aligned} \Delta_i &= ||\sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h}Ch^{\beta_i}e^{-\varepsilon_i mh}e^{(A^*+I/2)^T(n-m)h}|| \leq \\ &\leq C'\sum_{m=0}^{n} e^{-2\overline{\alpha}(n-m)h}\cdot h^{\beta_i}e^{-\varepsilon_i mh}. \end{aligned}$$

Applying Lemma 8.5 we get, assuming that $2\overline{\alpha} \neq \varepsilon_i$, with

$$\overline{\varepsilon}_i = \min(2\overline{\alpha}, \varepsilon_i) \tag{6.26}$$

the upper bound

$$\Delta_i \leq C' h^{\beta_i} e^{-\overline{\varepsilon}_i nh} / |e^{2\overline{\alpha}h} - e^{\varepsilon_i h}| = O(h^{\beta_i - 1} e^{-\overline{\varepsilon}_i nh}), \tag{6.27}$$

and thus

$$\begin{aligned} R^{\tilde{x}}_{3,n+1} &= \sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h}\ hP^*\ e^{(A^*+I/2)^T(n-m)h} + \\ &+ \sum_{m=0}^{n} O(h^{\beta_i - 1} e^{-\overline{\varepsilon}_i nh}). \end{aligned} \tag{6.28}$$

Obviously we have $\overline{\varepsilon}_i > 0$ for $i \neq 4$. For $i = 4$ we have $\overline{\varepsilon}_4 = 0$, but then $\beta_4 = 2$.

Next we consider the first, dominant term on the right hand side of (6.28) and define its approximation by setting $m' = n - m$ and extending the summation to ∞:

$$R^{\tilde{x}d}_{3,n+1} = \sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h} hP^* e^{(A^*+I/2)^T(n-m)h} \tag{6.29}$$

$$R^{\tilde{x}}_{3*} = \sum_{m'=0}^{\infty} e^{(A^*+I/2)m'h} hP^* e^{(A^*+I/2)^T m'h}. \tag{6.30}$$

Claim. *We have*

$$R^{\tilde{x}d}_{3,n+1} - R^{\tilde{x}}_{3*} = O(e^{-2\overline{\alpha}nh}). \tag{6.31}$$

Indeed, writing $m' = n - m$ and taking out the left-factor $e^{(A^*+I/2)(n+1)h}$ and the right factor $e^{(A^*+I/2)^T(n+1)h}$ we have

$$\begin{aligned} R^{\tilde{x}d}_{3,n+1} - R^{\tilde{x}}_{3*} = & \sum_{m'=n+1}^{\infty} e^{(A^*+I/2)m'h} hP^* e^{(A^*+I/2)^T m'h} \\ e^{(A^*+I/2)(n+1)h} & \left(\sum_{m=0}^{\infty} e^{(A^*+I/2)mh} hP^* e^{(A^*+I/2)^T mh} \right) e^{(A^*+I/2)^T(n+1)h}, \end{aligned}$$

the operator norm of which is obviously majorized by $C' e^{-2\overline{\alpha}nh}$, as claimed.

Lemma 6.6 *We have*

$$R^{\tilde{x}}_{3*} - S^* = O(h). \tag{6.32}$$

Proof: The covariance matrix $R^{\tilde{x}}_{3*}$ is the solution of the algebraic Lyapunov- equation

$$R^{\tilde{x}}_{3*} = e^{(A^*+I/2)h} R^{\tilde{x}}_{3*} e^{(A^*+I/2)^T h} + hP^*.$$

Taking into account the equality $e^{(A^*+I/2)h} = I + (A^* + I/2)h + O(h^2)$, this can be written as

$$R^{\tilde{x}}_{3*} = (I + (A^* + I/2)h + O(h^2)) R^{\tilde{x}}_{3*} (I + (A^* + I/2)^T h + O(h^2)) + hP^*$$

which is simplified to

$$0 = (A^* + I/2) R^{\tilde{x}}_{3*} + R^{\tilde{x}}_{3*} (A^* + I/2)^T + P^* + O(h),$$

and the stability of $(A^* + I/2)$ implies the claim. Combining (6.28), (6.31) and (6.32) we get, assuming that $2\overline{\alpha} \neq \varepsilon_i$,

$$R^{\tilde{x}}_{3,n+1} = S^* + O(e^{-2\overline{\alpha}nh} + h) + \sum_{i=1}^{5} O(h^{\beta_i - 1} e^{-\overline{\varepsilon}_i nh}). \tag{6.33}$$

The final approximation of $R^{\tilde{x}}_{3,n+1}$. For a given r we choose h and n in the following way: let $\varepsilon_h > 0$ and let h satisfy

$$e^{-\varepsilon_h r} \le h \le 2e^{-\varepsilon_h r}, \tag{6.34}$$

and in addition let r be an integer multiple of h, say $r = nh$. Then from (6.33) we get

$$R^{\tilde{x}}_{3,n+1} = S^* + O(e^{-2\overline{\alpha}nh} + e^{-\varepsilon_h nh}) + \sum_{i=1}^{5} O(e^{-(\beta_i - 1)\varepsilon_h nh} e^{-\overline{\varepsilon}_i nh}). \tag{6.35}$$

Combining this with (6.10) and substituting $h = e^{-\varepsilon_h r} = e^{-\varepsilon_h nh}$ we get

$$\begin{aligned} \mathrm{E}[\tilde{x}_{nh}\tilde{x}^T_{nh}] = S^* + O(e^{-2\overline{\alpha}nh} + e^{-\varepsilon_h nh}) + \sum_{i=1}^{5} O(e^{-(\beta_i - 1)\varepsilon_h nh} e^{-\overline{\varepsilon}_i nh}) + \\ O(e^{-\varepsilon_x nh} + e^{-\varepsilon_h nh/2} e^{-\varepsilon_{x2} nh} + e^{c\varepsilon_h nh} e^{-\varepsilon_{x3} nh}). \end{aligned} \tag{6.36}$$

The generic form of the error terms is $O(e^{-\gamma nh})$, where the values of γ are the following:

$$\begin{array}{lll} 2\overline{\alpha}, & \varepsilon_h, & (\beta_i - 1)\varepsilon_h + \overline{\varepsilon}_i,\ i = 1, ..., 5, \\ \varepsilon_x, & \varepsilon_h/2 + \varepsilon_{x2}, & c\varepsilon_h - \varepsilon_{x3}. \end{array}$$

Obviously for sufficiently small ε_h all these constant are positive and thus (6.8) and the claim of Theorem 6.2 follows.

7 Two applications

The usefulness of the results of the present paper is demonstrated by describing two applications. In the first example the pathwise cumulative regret is quantified for an on-line adaptive predictor of multi-variable linear stochastic systems, see (7.8). It is a previously unpublished result, presented at MTNS'96. In the second example a similar measure of performance degradation for the minimum-variance self-tuning regulator is considered. This problem, that had been formulated as far back as 1971 in [2] in a slightly different context from ours, has been solved only in 1994, see [28]. The result of [28] is restated in (7.19). A further application for indirect adaptive control of multi-variable linear stochastic systems is given in [27]. All these applications rely on the results of the present paper, in particular Theorems 4.3, 5.1. and 6.2.

Multivariable adaptive prediction. Let $(y_n), \quad 0 \le n < \infty$ be a vector-valued, wide-sense stationary stochastic process defined by a finite-dimensional linear stochastic system:

$$y = H(\theta^*)e. \tag{7.1}$$

Here $H(\theta) = I + C(\theta)(q^{-1}I - A(\theta))^{-1}B(\theta)$ is a square, causal, rational transfer function of the backward shift operator q^{-1}.

Condition 7.1 *$H(\theta)$ is defined for $\theta \in D$, where $D \subset \mathbb{R}^p$ is an open set and in its state-space realization the matrices $(A(\theta), B(\theta), C(\theta))$ are twice continuously differentiable functions of θ. Moreover $H(\theta)$ is stable and inverse stable.*

Condition 7.2 *The system-noise process $(e_n), \quad 0 \le n < \infty$ is an M-bounded, vector-valued wide-sense stationary orthogonal process. In addition there is an increasing sequence of σ-fields $(\mathcal{F}_n), \quad 0 \le n < \infty$, such that (e_n) is a martingale-difference process with constant conditional covariance:*

$$\mathrm{E}\,[e_n|\mathcal{F}_{n-1}] = 0, \qquad E(e_n e_n^T|\mathcal{F}_{n-1}) = \Lambda^*$$

almost surely, with $\Lambda^* > 0$.

These conditions will be called the standard conditions for multivariable linear stochastic systems. In the multivariable version of the prediction error method we have to estimate θ^* and Λ^* jointly

to improve efficiency. Let $\theta \in D$ and let Λ be a symmetric positive definite matrix and then define the second order stationary process $\overline{\varepsilon}(\theta)$ by

$$\overline{\varepsilon}(\theta) = H^{-1}(\theta)y.$$

Then define the cost function

$$V_N(\theta, \Lambda) = \frac{1}{2}\sum_{n=1}^{N} \overline{\varepsilon}_n^T(\theta)\Lambda^{-1}\overline{\varepsilon}_n(\theta) + \frac{N}{2}\log\ \det\Lambda. \tag{7.2}$$

If (e_n) is an i.i.d. sequence of Gaussian random vectors with distribution $N(0, \Lambda^*)$, then $V_N(\theta, \Lambda)$ is the negative conditional log-likelihood function, except for an additive constant. This cost function will be minimized in (θ_N, Λ_N) and the minimizing value, the off-line estimator of (θ^*, Λ^*) will be denoted by $(\widehat{\theta}_N, \hat{\Lambda}_N)$. A more precise definition of $(\widehat{\theta}_N, \hat{\Lambda}_N)$, taking into account the possibility of the existence of several local minima, can be given following [18].

Define the asymptotic cost function by

$$W(\theta, \Lambda) = \lim_{n\to\infty} \frac{1}{2}\mathrm{E}\ [\overline{\varepsilon}_n^T(\theta)\Lambda^{-1}\overline{\varepsilon}_n(\theta)] + \frac{1}{2}\log\ \ \det\Lambda. \tag{7.3}$$

It is easy to see that for any symmetric, positive definite Λ

$$W_\theta(\theta^*, \Lambda) = 0. \tag{7.4}$$

The Hessian of W with respect to θ at (θ^*, Λ^*) is

$$R^* = W_{\theta\theta}(\theta^*, \Lambda^*) = \lim_{n\to\infty} \mathrm{E}\ [\overline{\varepsilon}_{\theta n}^T(\theta^*)(\Lambda^*)^{-1}\overline{\varepsilon}_{\theta n}(\theta^*)]. \tag{7.5}$$

The above cost function can be treated with the extension of the DFL scheme indicated by an alternative definition of the random filed $H(n, x, \omega)$ in (3.47).

Condition 7.3 *The equation (7.4) has a unique solution $\theta = \theta^*$ for any symmetric, positive definite Λ and the Hessian-matrix $W_{\theta\theta}(\theta^*, \Lambda^*)$ is positive definite.*

The *performance index* of interest is the squared absolute value of the prediction error. Let $\Sigma_{\theta\theta}$ be the asymptotic covariance matrix of the off-line prediction error estimator $\widehat{\theta}_n$. Then it is well-known that $\Sigma_{\theta\theta} = (R^*)^{-1}$. Let

$$T^* = 2\frac{\partial^2}{\partial\theta^2} \lim_{n\to\infty} \mathrm{E}\ [\overline{\varepsilon}_n^T(\theta)\overline{\varepsilon}_n(\theta)]\ \big|_{\theta=\theta^*} \tag{7.6}$$

be the second-order sensitivity matrix of the performance index. Then we have the following result:

Theorem 7.1 *Let us consider a multivariable system satisfying Conditions 7.1, 7.2 and 7.3. In addition assume that (e_n) is L-mixing. Then we have almost surely*

$$\lim_{N\to\infty} \sum_{n=1}^{N} (|\overline{\varepsilon}_n(\widehat{\theta}_{n-1})|^2 - |e_n|^2)/\log N = \frac{1}{2}\mathrm{Tr}T^*\Sigma_{\theta\theta}. \tag{7.7}$$

The expression $\frac{1}{2}\mathrm{Tr}T^*$ will be called the *normalized cost of adaptation*. An important difference between ARMA and multivariable systems is that, unless $\Lambda^* \neq cI$, with c being a scalar, the trace-formula given on the right hand side of (7.7) can not be further simplified. However it can be shown that $\frac{1}{2}\mathrm{Tr}T^*\Sigma_{\theta\theta}$ is invariant with respect to diffeomorphic transformation of the parameter-space, while restriction of the parameter space, i.e. writing $\theta = g(\eta)$ with $\dim\ \eta < \dim\ \theta$, with g being a smooth function, reduces the normalized cost of adaptation. Note, that the normalized cost of adaptation is not determined solely by structural parameters, it may depend also on the actual multivariable system, unlike in the ARMA case.

To extend this result for adaptive predictors defined in terms of recursive estimators we rely on Theorem 4.3 and we get the following result:

Claim. *Let $\widehat{\widehat{\theta}}_n$ be a recursive estimator of θ^* with asymptotic covariance matrix $\overline{\Sigma}_{\theta\theta}$. Then under appropriate technical conditions, obtained by specializing the conditions of Theorem 4.3, we have*

$$\lim_{N\to\infty}\sum_{n=1}^{N}(|\overline{\varepsilon}_n(\widehat{\widehat{\theta}}_{n-1})|^2-|e_n|^2)/\log N=\frac{1}{2}\mathrm{Tr}T^*\overline{\Sigma}_{\theta\theta} \tag{7.8}$$

almost surely. In analogy with the ARMA-case, if we use a stochastic Newton method, then we have

$$\overline{\Sigma}_{\theta\theta}=\Sigma_{\theta\theta}.$$

The minimum-variance self-tuning regulator. Consider now a stochastic control system in ARMAX(n,m,p) representation defined by the relation

$$A^*(q^{-1})y=q^{-1}B^*(q^{-1})u+C^*(q^{-1})e, \tag{7.9}$$

where $A^*(q^{-1}), B^*(q^{-1})$ and $C^*(q^{-1})$ are polynomials of the backward shift operator q^{-1} of degree n,m,p respectively. Their coefficients are denoted by a_i^*, b_i^*, c_i^*, respectively, with $a_0^*=1, a_n^*\neq 0$, $b_0^*\neq 0, b_m^*\neq 0$, $c_0^*=1, c_p^*\neq 0$. Here u is the input process, e is the noise process and y is the output process. The notation u is a shorthand for $(u(t)), 0\leq t\leq\infty$. Assume that the polynomials B^* and C^* are stable and that $\deg C^*\leq\deg A^*$. By extending the vector $(c_1^*,\dots,c_p^*)$ with zeros, if necessary, we can actually assume that $\deg A^*=\deg C^*$. The stochastic process e is a zero mean wide-sense stationary orthogonal process, i.e. for all $t,s\geq 0$ we have $\mathrm{E}e(t)=0$ and $\mathrm{E}\,[e(s)e(t)]=\sigma^2(e)\delta_{st}$, where δ_{st} is the Kronecker-symbol.

The minimum-variance control for the ARMAX system given under (7.9) is given by (cf. [2])

$$q^{-1}B^*u=(A^*-C^*)y. \tag{7.10}$$

Using this control law we get, under the assumption that the initial values are all zero, $y(t)=e(t)$. (7.10) can be written in the form

$$u(t-1)=-(\eta^*)^T\phi(t), \tag{7.11}$$

where

$$\eta^*=\frac{1}{b_0^*}(a_1^*-c_1^*,\dots,a_n^*-c_n^*,b_1^*,\dots b_m^*)^T \tag{7.12}$$

and

$$\phi(t)=(-y(t-1),\dots,-y(t-n),u(t-2),\dots,u(t-m-1)). \tag{7.13}$$

If the values of the parameters of the stochastic control system are unknown then a stochastic adaptive control procedure will be needed. Within stochastic adaptive control a special procedure is the self-tuning regulation, that has been proposed in [2] for minimum variance control. For a new perspective of this procedure see [63]. This is a stochastic approximation procedure defined as follows: let $\hat{\eta}(0)$ be an initial estimate of η^* and let $\hat{\eta}(t-1)$ be an estimate computed at time $t-1$. Then define the control action by

$$u(t-1)=-\hat{\eta}(t-1)\ \phi(t). \tag{7.14}$$

This is followed by observing $y(t)$ which is generated by (7.9). Finally we generate the next estimates $\hat{\eta}(t)$ by

$$\hat{\eta}(t)=\hat{\eta}(t-1)+R^{-1}\frac{1}{t}\ \phi(t)y(t), \tag{7.15}$$

where R is a symmetric positive definite matrix. A basic question in the context of stochastic adaptive control is the characterization of the performance degradation

$$y^2(t)-e^2(t) \tag{7.16}$$

and to establish its pathwise properties. This problem was first formulated in [2]. It has been open for a long time, until a solution was presented in [28], using the results of the present paper.

The performance of the minimum variance self-tuning regulator had been studied in [48, 49]. In [48] the right order of magnitude for the so-called cumulative regret was found for general ARMAX systems. In [49] the right constant in a tight upper bound for cumulative regret had been obtained for ARX systems. For a survey see [46]. Note, however, that in these papers so-called indirect adaptive control procedures had been considered, where identifiability is ensured by the injection of rare shocks with diminishing frequency into the system. Similar results were obtained in [31] and in [32].

Let $D \subset \mathbb{R}^{n+m}$ be a set of candidate controller-parameters to be specified below. For any $\eta \in D$ and for $t \geq 0$ we consider the control law

$$u(t-1) = -\eta^T \phi(t),$$

where $\phi(t)$ is defined above in (7.13). Thus we get a closed-loop system in which both u and y depend on η. To stress this dependence we write $u(t) = \bar{u}(t,\eta)$ and $y(t) = \bar{y}(t,\eta)$. Let D denote the open set of η's in $\mathbb{R}^{m+n}$ such that the closed loop system is stable. Define the nonlinear vector-valued function

$$G(\eta) \stackrel{\Delta}{=} \lim_{t\to\infty} \mathrm{E}\,[\,\bar{\phi}(t,\eta)\bar{y}(t,\eta)]. \tag{7.17}$$

It is easy to see that we have $G(\eta^*) = 0$. Let S^* denote the asymptotic covariance matrix of $\hat{\eta}(t)$ i.e. let

$$S^* = \lim_{t\to\infty} t \cdot \mathrm{E}\,[(\hat{\eta}(t) - \eta^*)(\hat{\eta}(t) - \eta^*)^T],$$

assuming that the limit exists. Define the *second order sensitivity matrix*

$$T^* = \lim_{t\to\infty} \mathrm{E}\,[\frac{\partial^2}{\partial \eta^2}_{|\eta=\eta^*} \bar{y}^2(t,\eta)]. \tag{7.18}$$

Claim (see [28]). *Consider the minimum-variance self-tuning regulator for an ARMAX(n,m,p) system given by (7.15). Then, under appropriate technical conditions, obtained by specializing the conditions of Theorem 4.3, we have the following pathwise characterization of the cumulative performance degradation:*

$$\lim_{N\to\infty} \sum_{t=1}^{N} (y^2(t) - e^2(t))/\log N = \frac{1}{2}\mathrm{Tr}\; T^* S^* \tag{7.19}$$

almost surely. Moreover, for any symmetric positive definite R we have

$$\frac{1}{2}\mathrm{Tr}\; T^* S^* \geq \sigma^2(e)(m+n). \tag{7.20}$$

The inequality (7.20) is an equality if and only if $R = -G_\eta(\eta^*)$ and $C^* = 1$.

The proof of (7.19) follows [24]. We note in passing that it has been a common belief that $G_\eta(\eta^*)$ is not computable. However, using a technique of Hjalmarsson (cf. [39]) it can be shown that for certain interesting physical systems $G_\eta(\eta^*)$ is in fact computable. The proof of (7.19) follows [24].

Conclusion. Performance degradation due to statistical uncertainty, also called regret, is of great interest in adaptive prediction and control of stochastic systems. To quantify the pathwise cumulative regret we need technical tools similar to those developed in [24] in the context of adaptive prediction of ARMA-processes. These new tools have been developed in this paper. The usefulness of the results in stochastic adaptive control has been demonstrated for the minimum-variance self-tuning regulator for ARMAX-systems in Section 7, see also [28]. A further application for indirect adaptive control of multi-variable linear stochastic systems is given in [27].

The results can be also applied in the context of identification for control, see [29, 40, 41]. For any fixed feedback strategy the covariance matrix of the estimation error and consequently the

cumulative regret over any finite horizon will depend on the feedback strategy. The cumulative regret over finite horizon distorts the performance of the controller and this distortion can be precisely characterized using the results of the present paper. Thus a controller with optimal overall performance over a fixed finite horizon can be developed, at least in theory, i.e. pretending that we know the systems dynamics.

A further potential area of application is adaptive experimental design, see [30], in which the objective function to be minimized is the trace of the covariance matrix of the estimation error, which can be computed experimentally for any fixed input pattern.

Another, more classical possible application is the derivation of limit results such as LIL and invariance principles along the lines of [38].

The scope of applications can be enlarged by extending the technical results themselves. The extension of the results of the present paper to Kiefer-Wolfwitz-type stochastic approximation procedures, such as the simultaneous perturbation stochastic approximation or SPSA method due to Spall, [61] and [62], seems to be possible.

8 Appendix: Auxiliary results

Lemma 8.1 *Let $(x_t), t \geq 0$ be a zero-mean L-mixing process with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$ and let y be an $\mathcal{F}_s$-measurable random variable for some $0 \leq s \leq t$, such that its moments, which appear in the inequality below, are finite. Then for every $1 \leq p, q \leq \infty$ such that $1/p + 1/q = 1$ we have*

$$|\mathrm{E} x_t y| \leq 2\gamma_p(t-s, x)\mathrm{E}^{1/q}|y|^q.$$

Analogous inequalities for strong mixing stationary sequences are given in [9] and for uniformly mixing stationary sequences in [42]. A concise survey of these inequalities is given in Chapter 7.2 of [15] and Appendix III of [33]. Here we restate an improved Hölder inequality under the weakest condition on mixing, namely strong-mixing or α-mixing (cf. Corollary 2.5 of Chapter 7.2 of [15]):

Lemma 8.2 *Let $p, q, r > 1$ be such that $p^{-1} + q^{-1} + r^{-1} = 1$. Let Y and Z be $\mathcal{H}$-measurable and $\mathcal{G}$-measurable random variables such that $||Y||_q$ and $||Z||_r$ are finite, respectively. Then*

$$|\mathrm{E}[YZ] - \mathrm{E}[Y]\mathrm{E}[Z]| \leq C\alpha(\mathcal{H}, \mathcal{G})^{1/p}||Y||_q||Z||_r. \tag{8.1}$$

The improved Hölder inequality of Lemma 8.1 plays a key role in deriving the following moment-inequality (cf. Theorem 1.1 in [17]):

Theorem 8.1 *Let $(u_t), t \geq 0$ be a zero-mean L-mixing process. Let (f_t) be a function in $L_2[0, T]$. Then we have for all $m \geq 2$ with $C_m = 2(m-1)^{1/2}$*

$$\mathrm{E}^{1/m}|\int_0^T f_s u_s ds|^m \leq C_m(\int_0^T f_t^2 dt)^{1/2} M_m^{1/2}(u) \cdot \Gamma_m^{1/2}(u).$$

Extension of the statement to vector-valued processes is an elementary exercise, but obviously the constant C_m will be different. Extension to random (f_t) is not possible in general, but an extension is possible for multiple integrals with deterministic kernel (cf. [21]). Here we need only the following special result:

Theorem 8.2 *Let (u_t) and (v_t) be zero mean L-mixing processes. Then we have*

$$I_{T_0} = \int_{T_0}^T \frac{1}{t} u_t \int_{T_0}^t \frac{1}{s} v_s ds dt = O_M(T_0^{-1}).$$

The following simple lemma is stated as Corollary 3.5 in [24].

Lemma 8.3 *Let $(\mathcal{F}_t, \mathcal{F}_t^+)$ be a pair of families of σ-algebras as in Section 3 and let $(x_t), t \geq 0$ be an $\mathcal{F}_t$-adapted, measurable stochastic process. Then (x_t) is L-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$ if and only if the processes (x_{n+d}) are L-mixing with respect to $(\mathcal{F}_{n+d}, \mathcal{F}_{n+d}^+)$, uniformly in d for $0 \leq d < 1$.*

Let us consider a stochastic process $(u_n(\theta))$ with $\theta \in D \subset \mathbb{R}^p$, where D is an open set, which is measurable, separable, M-bounded and M-Lipschitz-continuous in θ for $\theta \in D$. By Kolmogorov's theorem the realizations of $(x_n(\theta))$ are continuous in θ with probability 1, hence we can define for almost all ω

$$u_n^* = \max_{\theta \in D_0} |u_n(\theta)|,$$

where $D_0 \subset D$ is a compact domain. The following result is given as Theorem 3.4 in [17].

Theorem 8.3 *Assume that $(u_n(\theta))$ is a stochastic process which is measurable, separable, M-bounded and M-Lipschitz continuous in θ for $\theta \in D$. Let u_n^* be the random variable defined above. Then we have for all positive integers q and $s > p$*

$$M_q(u^*) \leq C(M_{qs}(u) + M_{qs}(\Delta u/\Delta\theta)),$$

where C depends only on $p, q, s,$ and D_0, D.

A continuous-time version of following lemma was given in [17] as Lemma 2.4:

Lemma 8.4 *Let $(u_n), n \geq 0$ be a zero-mean L-mixing $\mathbb{R}^p$-valued process and define another $\mathbb{R}^p$-valued process (x_n) by*

$$x_{n+1} = Ax_n + u_n, \qquad x_0 = 0,$$

where the spectral norm of A is smaller than 1, say we have $||A^n|| \leq C\alpha^n$ with some $C > 0$ and $0 < \alpha < 1$. Then the output process (x_n) is L-mixing.

The first part of the following result was stated in Lemma 7.4 of [19]. The second part of the quoted lemma was not correctly stated and is therefore restated and proved here.

Lemma 8.5 *Let $(u_n), n \geq 0$ be an M-bounded process and define a process (x_n) by*

$$x_{n+1} = \lambda x_n + \rho^n u_n, \qquad x_0 = 0, \tag{8.2}$$

where $0 < \lambda < \rho$. Then for any $m \geq 1$ we have

$$\mathrm{E}^{1/m}|x_n|^m \leq \frac{\rho^n}{\rho - \lambda} M_m(u).$$

On the other hand if $0 < \rho < \lambda$ then we have

$$\mathrm{E}^{1/m}|x_n|^m \leq \frac{\lambda^n}{\lambda - \rho} M_m(u).$$

Proof: Let $0 < \lambda < \rho$ and set $z_n = \rho^{-n} x_n$. Then we have, after multiplying (8.2) by $\rho^{-(n+1)}$,

$$z_{n+1} = \lambda\rho^{-1} z_n + \rho^{-1} u_n,$$

which can be solved explicitly for z_n to get

$$z_n = \sum_{i=0}^{n-1} (\lambda\rho^{-1})^i \rho^{-1} u_{n-1-i}.$$

Using the triangle inequality for the $L_m(\Omega, \mathcal{F}, P)$ norm and the condition $0 < \lambda < \rho$ we get

$$M_m(z) \leq (1 - \lambda\rho^{-1})^{-1} \rho^{-1} M_m(u)$$

from which the first proposition follows.

A useful reformulation of the above argument is the following: writing

$$x_n = \rho^n z_n = \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} u_{n-1-i} \tag{8.3}$$

we have

$$\mathrm{E}^{1/m}|x_n|^m \le \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} \mathrm{E}^{1/m}|u_{n-1-i}|^m \le \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} M_q(u). \tag{8.4}$$

Thus it is sufficient to establish that for $0 < \lambda < \rho$

$$\sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} \le \frac{\rho^n}{\rho - \lambda} \tag{8.5}$$

and this is obtained from the above argument with $u_n = 1$ for all n. The advantage of this reformulation is that the left hand side is the convolution of the sequences (λ^n) and (ρ^n) and thus it is symmetric in λ and ρ.

In the case when $0 < \rho < \lambda$ we use (8.4) to estimate $\mathrm{E}^{1/m}|x_n|^m$, but the role of λ and ρ is interchanged, thus we get

$$\mathrm{E}^{1/m}|x_n|^m \le \frac{\lambda^n}{\lambda - \rho} M_m(u).$$

Remark. A simple corollary is that

$$\sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} \le \frac{\max\ (\lambda^n, \rho^n)}{|\rho - \lambda|}. \tag{8.6}$$

The lemma below has been used for ODE analysis of stochastic approximation processes in [16]. The conditions are similar to Condition 3.4, (i). Consider the ordinary differential equation

$$\dot{y}_t = F(t, y_t), \qquad y_s = \xi,\ s \ge 1. \tag{8.7}$$

The solution of the above ODE will be denoted by $y(t, s, \xi)$ in the time interval where it exists and is unique.

Condition 8.1 *$F = (F(t, y))$ is defined for $t \ge 1, y \in D$ where $D \subset \mathbb{R}^p$ is an open set and F is continuously differentiable in (t, y). It is assumed that there exists a compact domain $D_0' \subset D$ such that $y(t, s, \xi) \in D$ for all $\xi \in D_0'$ and $1 \le s \le t < \infty$.*

Lemma 8.6 *Assume that Condition 8.1 is satisfied. Let $(x_t), 1 \le t < \infty$ be a continuous, piecewise continuously differentiable curve such that $x_t \in D_0'$ for $t \ge 1$ and $x_1 = y_1 = \xi \in D_0'$. Then for $t \ge 1$*

$$x_t - y_t = \int_1^t \frac{\partial}{\partial \xi} y(t, r, x_r) \Big(\dot{x}_r - F(r, x_r) \Big)\ dr. \tag{8.8}$$

Proof: Write $z_r = y(t, r, x_r)$. Obviously the left hand side of (8.8) can be written as $z_t - z_1$ and we have

$$z_t - z_1 = \int_1^t \dot{z}_r dr = \int_1^t \Big(y_r(t, r, x_r) + y_\xi(t, r, x_r) \dot{x}_r \Big)\ dr. \tag{8.9}$$

Taking into account the equality $y_r(t, r, x_r) = -y_\xi(t, r, x_r) \cdot F(t, x_r)$ we get the lemma.

A discretized version of the above lemma has been used implicitly in the final step of the proof of Theorem 1.1. of [19], see (2.10) of [19]. We now formulate this lemma with explicit conditions. It has been used in the proof of Lemma 4.4.

Condition 8.2 *Assume that D_0' is convex and there exists a compact set $D_0 \subset D_0'$ such that for all $x \in D_0$ and $t \ge s \ge 1$ we have $y(t, s, x) \in D_0'$.*

Let $1 = s_0 \le s_2 \le ... \le s_n \le s_{n+1} = t$ and let $(x_{s_i}) \in D_0,\ i = 0, 1, ..., n$ be a sequence such that $x_1 = y_1 = \xi \in D_0$. These point are considered as approximations to $y_{s_i} = y(s_i, 1, \xi)$. We will estimate the tracking error $x_t - y_t$ in terms of *local tracking* errors

$$(x_{s_i} - y(s_i, s_{i-1}, x_{s_{i-1}})).$$

Lemma 8.7 *Let $F = (F(t,y))$ satisfy Conditions 8.1 and 8.2 and let $(x_{s_i}) \in D_0,\ i = 0,1,...,n$ be a sequence such that $x_1 = y_1 = \xi$. Then*

$$x_t - y_t = (x_t - y(t, s_n, x_{s_n})) + \sum_{i=1}^{n} \int_0^1 \frac{\partial}{\partial \xi} y(t, s_i, w(i,\lambda))\ d\lambda \cdot (x_{s_i} - y(s_i, s_{i-1}, x_{s_{i-1}})), \quad (8.10)$$

where $w(i,\lambda) = (1-\lambda)y(s_i, s_{i-1}, x_{s_{i-1}}) + \lambda x_{s_i}$.

Proof: Consider the sequence $z_i = y(t, s_i, x_{s_i}),\ i = 0,1,...,n$. Then $z_0 = y_t$ and we can write

$$x_t - y_t = (x_t - z_n) + \sum_{i=1}^{n} (z_i - z_{i-1}) = (x_t - y(t, s_n, x_{s_n})) + \sum_{i=1}^{n} (y(t, s_i, x_{s_i}) - y(t, s_{i-1}, x_{s_{i-1}})). \quad (8.11)$$

Now for $1 \le s \le s' \le t$ we have $y(t,s,x) = y(t,s',y(s',s,x))$. Setting $s = s_{i-1}, s' = s_i, x = x_{s_{i-1}}$ the i-th term of the right hand side of (8.11) thus becomes

$$y(t, s_i, x_{s_i}) - y(t, s_i, y(s_i, s_{i-1}, x_{s_{i-1}})) = \int_0^1 \frac{\partial}{\partial \xi} y(t, s_i, w(i,\lambda))\ d\lambda \cdot (x_{s_i} - y(s_i, s_{i-1}, x_{s_{i-1}}))$$

with $w(i,\lambda) = (1-\lambda)y(s_i, s_{i-1}, x_{s_{i-1}}) + \lambda x_{s_i}$ for $0 \le \lambda \le 1$. Note that $w(i,\lambda) \in D_0'$ for $i = 1,...,n$ since D_0' is convex and thus $y(t, s_i, w(i,\lambda))$ is well-defined, and the lemma follows.

Let $G = (G(y))$ be defined in an open set $D \subset \mathbb{R}^p$ and consider the ordinary differential equation

$$\dot{y}_t = \frac{1}{t} G(y_t), \qquad y_s = \xi,\ s \ge 1. \quad (8.12)$$

We will have conditions that ensure that the above ODE has a unique solution in some finite or infinite interval, which we denote by $y(t,s,\xi)$. We assume the validity of the following condition, which is weaker than Conditions 3.3 and 3.4.

Condition 8.3 *G has continuous partial derivatives up to second order for $y \in D$. There exists compact sets $D_0 \subset D_0' \subset D$ such that for all $\xi \in D_0,\ t \ge s \ge 1$ we have $y(t,s,\xi) \in D_0'$ and*

$$\|y_\xi(t,s,\xi)\| \le C_0 (s/t)^\alpha \quad (8.13)$$

with some $C_0 \ge 1, \alpha > 0$. Let $||\partial^i G(y)/\partial y^i|| \le L$ for $y \in D_0'$ and $i = 0,1,2$.

We prove that the stability expressed by the condition above is in a sense inherited by the second order derivatives of $y(t,s,\xi)$.

Lemma 8.8 *Let G satisfy Condition 8.3. Then for all $\xi \in D_0,\ t \ge s \ge 1$*

$$\begin{aligned} \|y_{\xi\xi}(t,s,\xi)\| &\le L\alpha^{-1} C_0^3 \cdot (s/t)^\alpha, \\ \|y_{s\xi}(t,s,\xi)\| &\le (L\alpha^{-1} + 1) L C_0^3 \cdot \frac{1}{s} (s/t)^\alpha. \end{aligned}$$

Remark. From the proof below it follows that if G is three-times continuously differentiable then with some constant C_0' we have $\|y_{\xi\xi\xi}(t,s,\xi)\| \le C_0'(s/t)^\alpha$.

Proof: Use a change of time-scale $t = e^v,\ s = e^u$ and consider the differential equation

$$\frac{d}{dv} z_v = G(z_v), \qquad z_u = \xi,\ u \ge 0,$$

with its solution being denoted by $z(v,u,\xi),\ v \ge u \ge 0$. Then (8.13) implies

$$\|z_\xi(v,u,\xi)\| \le C_0 e^{-\alpha(u-v)}, \quad (8.14)$$

and the propositions of the lemma is equivalent to, after substitution $u = \log t$ and $v = \log s$,

$$\begin{aligned} \|z_{\xi\xi}(v,u,\xi)\| &\le L\alpha^{-1} C_0^3 \cdot e^{-\alpha(v-u)}, \\ \|z_{u\xi}(v,u,\xi)\| &\le (L\alpha^{-1} + 1) L C_0^3 \cdot e^{-\alpha(v-u)}. \end{aligned}$$

Now we have

$$\frac{\partial}{\partial v} z_\xi(v,u,\xi) = G_y(z(v,u,\xi)) \cdot z_\xi(v,u,\xi), \qquad z_\xi(u,u,\xi) = I. \tag{8.15}$$

It is easy to see that $z_{\xi\xi}(v,u,\xi)$ exists and is continuous in (v,u,ξ). From (8.15) get

$$\frac{\partial}{\partial v} z_{\xi\xi}(v,u,\xi) = G_{yy}(z(v,u,\xi)) \cdot z_\xi(v,u,\xi) z_\xi(v,u,\xi) + + G_y(z(v,u,\xi)) \cdot z_{\xi\xi}(v,u,\xi), \tag{8.16}$$

with $z_{\xi\xi}(u,u,\xi) = 0$. Since the operator norm of the first term is majorized by $LC_0^2 e^{-2\alpha(u-v)}$ and since the time varying linear differential equation with transition matrix $G_y(z(v,u,\xi))$ is exponentially stable due to (8.14), we get the first claim of the lemma from the identity

$$\int_0^t e^{-\alpha(v-r)} e^{-2\alpha r} dr = e^{-\alpha v} \int_0^v e^{-\alpha r} dr < \alpha^{-1} e^{-\alpha v}.$$

To estimate the mixed derivatives, take into account $z_u(v,u,\xi) = -z_\xi(v,u,\xi) \cdot G(\xi)$ to get

$$z_{u\xi}(v,u,\xi) = -z_{\xi\xi}(v,u,\xi) \cdot G(\xi) - z_\xi(v,u,\xi) \cdot G_\xi(\xi),$$

from which the second claim follows using (8.14) and the proven first part of the lemma.

Acknowledgement

This research was supported by the Natural Sciences and Engineering Research Council of Canada under Grant 01329 and by the National Research Foundation of Hungary under Grant T 047193. The author expresses his thanks to Peter Caines for arranging a visit to McGill University and for numerous discussions on linear stochastic systems and to Zalán Mátyás, Zsanett Orlovits and Zsuzsanna Vágó for carefully reading this paper.

References

[1] K.J. Åström and T. Söderström. Uniqueness of the maximum-likelihood estimates of the parameters of an ARMA modell. *IEEE Trans. Automat. Contr.*, AC-19:769–773, 1974.

[2] K.J. Åström and B. Wittenmark. Problems of identification and control. *J. Math. Anal. Appl.*, 34:90–113, 1971.

[3] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*. Springer-Verlag, Berlin, 1990.

[4] B. Bercu. Weighted estimation and tracking for ARMAX models. *SIAM J. Control and Optimization*, 33:89–106, 1995.

[5] A.N. Borodin. A stochastic approximation procedure in the case of weakly dependent observations. *Theory of Probability and Appl.*, 24:34–52, 1979.

[6] P.E. Caines. *Linear Stochastic Systems*. Wiley, 1988.

[7] H.F. Chen and L.Guo. *Identification and Stochastic Adaptive Control*. Birkhauser, 1991.

[8] L.D. Davisson. Prediction error of stationary Gaussian time series of unknown covariance. *IEEE Trans. Informat. Theory*, IT-19:783–795, 1965.

[9] Yu. A. Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory of Probab. Appl.*, 13:691–696, 1968.

[10] B. Delyon. General results on the convergence of stochastic algorithms. *IEEE Trans. Automat. Contr.*, 41:1245–1255, 1996.

[11] D.P. Djereveckii and A.L. Fradkov. Application of the theory of Markov-processes to the analysis of the dynamics of adaptation algorithms. *Automation and Remote Control*, (2):39–48, 1974.

[12] D.P. Djereveckii and A.L. Fradkov. Two models for analysing the dynamics of adaptation technics. *Automation and Remote Control*, (1):67–75, 1974.

[13] D.P. Djereveckii and A.L. Fradkov. *Applied theory of discrete adaptive control systems*. Nauka, Moscow, 1981. In Russian.

[14] T.E. Duncan and B. Pasik-Duncan. Some methods for the adaptive control of continuous time linear stochastic systems. In L. Gerencsér and P.E. Caines, editors, *Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control*, pages 242–267. Springer-Verlag Berlin, Heidelberg, 1991.

[15] S.N. Ethier and T.G.Kurtz. *Markov Processes. Characterization and Convergence*. Wiley, 1986.

[16] S. Geman. Some averaging and stability results for random differential equations. *SIAM Journal of Applied Mathematics*, 36:87–105, 1979.

[17] L. Gerencsér. On a class of mixing processes. *Stochastics*, 26:165–191, 1989.

[18] L. Gerencsér. On the martingale approximation of the estimation error of ARMA parameters. *Systems & Control Letters*, 15:417–423, 1990.

[19] L. Gerencsér. Rate of convergence of recursive estimators. *SIAM J. Control and Optimization*, 30(5):1200–1227, 1992.

[20] L. Gerencsér. A representation theorem for the error of recursive estimators. In *Proc. of the 31st IEEE Conference on Decision and Control, Tucson*, pages 2251–2256, 1992.

[21] L. Gerencsér. Multiple integrals with respect to L-mixing processes. *Statistics and Probability Letters*, 17:73–83, 1993.

[22] L. Gerencsér. Strong approximation of the recursive prediction error estimator of the parameters of an ARMA process. *Systems & Control Letters*, 21:347–351, 1993.

[23] L. Gerencsér. Fixed gain off-line estimators of ARMA parameters. *Journal of Mathematical Systems, Estimation and Control*, 4(2):249–252, 1994. Retreival code for full electronic manuscript: 66945.

[24] L. Gerencsér. On Rissanen's predictive stochastic complexity for stationary ARMA processes. *Statistical Planning and Inference*, 41:303–325, 1994.

[25] L. Gerencsér. Stability of random iterative mappings. In M. Dror, P. L'ecuyer, and F. Szidarovszky, editors, *Modeling Uncertainty. An Examination of its Theory, Methods, and Applications*, pages 359–371. Dordrecht, Kluwer, 2002.

[26] L. Gerencsér and J. Rissanen. A prediction bound for Gaussian ARMA processes. *Proc. of the 25th Conference on Decision and Control, Athens*, 3:1487–1490, 1986.

[27] L. Gerencsér and Zs. Vágó. Adaptive control of multivariable linear stochastic systems. a strong approximation approach. In *Proceedings of the European Control Conference, ECC'99, Karlsuhe, Germany*, page F587, 1999.

[28] L. Gerencsér, J.H. van Schuppen, J. Rissanen, and Zs. Vágó. Stochastic complexity, self-tuning and optimality. In *Proceedings of the 33rd CDC, Florida*, pages 652–654, 1994.

[29] M. Gevers. Towards a joint design of identification and control ? In H.L. Trentelman and J.C. Willems, editors, *Essays on Control : Perspectives in the Theory and its Applications*, pages 111–151. Birkhäuser, 1993.

[30] M. Gevers, X. Bombois, B. Codrons, F. De Bruyne, and G. Scorletti. The role of experimental conditions in model validation for control. In A. Garulli, A. Tesi, and A. Vicino, editors, *Robustness in Identification and Control*, pages 72–86. Lecture Notes in Control and Information Sciences, Vol. 245, Springer Verlag, 1999.

[31] L. Guo. The logarithm law of self-tuning regulators. In *Proceedings of the 12th IFAC World Congress, Sydney*, volume I, pages 227–232, 1993.

[32] L. Guo. Further results on least squares based adaptive minimum variance control. *SIAM J. on Control and Optimization*, 32(1):187–212, 1994.

[33] P. Hall and C.C. Heyde. *Martingale limit theory and its applications*. Academic Press, 1980.

[34] E.J. Hannan. The convergence of some time-series recursions. *Ann. Stat.*, 4:1258 – 1270, 1976.

[35] E.J. Hannan and M. Deistler. *The statistical theory of linear systems*. Wiley, 1988.

[36] E.J. Hannan, A.J. McDougall, and D.S. Poskitt. Recursive estimation of autoregressions. *J. Roy. Stat. Soc., Ser B.*, 51:217– 233, 1989.

[37] E.M. Hemerley and M.A.H. Davis. Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Stat.*, 17:941 – 946, 1989.

[38] A. Heunis. Rates of convergence for an adaptive filtering algorithm driven by stationary dependent data. *SIAM J. on Control and Optimization*, 32:116–139, 1994.

[39] H. Hjalmarsson. Efficient tuning of linear multivariable controllers using iterative feedback tuning. *Int. J. Adapt.Control Signal Process.*, 13:553–572, 1999.

[40] H. Hjalmarsson, M. Gevers, and F. De Bruyne. For model based control design criteria, closed loop identification gives better performance. *Automatica*, 32:1659–1673, 1996.

[41] H. Hjalmarsson and K. Lindqvist. Identification for control:L_2 and L_∞ methods. In *Conference on Decision and Control*, Orlando, Florida, USA, December 2001. IEEE.

[42] I.A. Ibragimov and Yu. A. Linnik. *Independent and stationary sequences of random variables*. Wolters and Nordhoff, Groningen, 1971.

[43] J. A. Joslin and A. J. Heunis. Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm. *SIAM J. on Control and Optimization*, 39:533–570, 2000.

[44] T. Kailath. *Linear Systems*. Prentice–Hall, 1980.

[45] H.J. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer Verlag. New York, 1997.

[46] T.Z. Lai. Information bounds, certainty equivalence and learning in asymptotically efficient adaptive control of time-invariant stochastic systems. In L. Gerencsér and P.E. Caines, editors, *Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control*, pages 268–299. Springer-Verlag Berlin, Heidelberg, 1991.

[47] T.Z. Lai and C.Z. Wei. Least squares estimates in stochastic regression models with application to identification and control of dynamic systems. *Ann. of Stat.*, 10:154 – 165, 1982.

[48] T.Z. Lai and C.Z. Wei. Extended Least Squares and their applications to adaptive control of dynamic systems. *IEEE Trans. Automat. Contr.*, 31:898–906, 1986.

[49] T.Z. Lai and C.Z. Wei. Asymptotically efficient self-tuning regulators. *SIAM J. Control and Optimization*, 25:466–481, 1987.

[50] L. Ljung. On consistency and identifiability. *Mathematical Programming Study*, 5:169–190, 1976.

[51] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Contr.*, 22:551–575, 1977.

[52] L. Ljung, G. Pflug, and H. Walk. *Stochastic Approximation and Optimization of Random Systems*. Birkhauser Verlag, DMV Seminar, Band 17, 1992.

[53] L. Ljung and T. Söderström. *Theory and practice of recursive identification*. The MIT Press, 1983.

[54] M.B. Nevel'son and R.Z. Has'minskii. *Stochastic Approximation and Recursive Estimation*. American Mathematical Soc., Providence RI, 1976.

[55] R. Ober. Balanced realizations: canonical form, parametrization, model reduction. *Internat. J. Control*, 46:643–670, 1987.

[56] J. Rissanen. A predictive least squares principle. *IMA Journal of Mathematical Control an Information*, 3(2-3):211–222, 1986.

[57] J. Rissanen. Stochastic complexity and predictive modelling. *Annals of Statistics*, 14(3):1080–1100, 1986.

[58] J. Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific Publisher, 1989.

[59] J. Rissanen and P.E. Caines. The strong consistency of maximum likelihood estimators for ARMA processes. *Ann. Statist.*, 7:297 – 315, 1979.

[60] V. Solo. The second order properties of a time series recursion. *Ann. Stat.*, 9:307–317, 1981.

[61] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Contr.*, 37:332–341, 1992.

[62] J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. In *Proceedings of the 1998 IEEE CDC*, pages 3872 – 3879, 1998.

[63] J.H. van Schuppen. Tuning of Gaussian stochastic control systems. Report BS-R9223, CWI, Amsterdam, 1992.

[64] S.M. Veres. Relations between information criteria for model-structure selection Part 3. Strong consistency of the predictive least squares criterion. *International Journal of Control*, 52:737–751, 1990.

[65] G. Yin. A stopping rule for the Robbins-Monro method. *J. of Optimization Theory and its Applications*, 67(1):151–173, 1990.