Convergence Rate of Moments in Stochastic Approximation with Simultaneous Perturbation Gradient Approximation and Resetting

László Gerencsér

***Abstract*— The sequence of recursive estimators for function minimization generated by Spall's simultaneous perturbation stochastic approximation (SPSA) method, presented in [25], combined with a suitable restarting mechanism is considered. It is proved that this sequence converges under certain conditions with rate $O(n^{-\beta/2})$ for some $\beta>0$, the best value being $\beta = 2/3$, where the rate is measured by the L_q-norm of the estimation error for any $1 \leq q < \infty$. The authors also present higher order SPSA methods. It is shown that the error exponent $\beta/2$ can be arbitrarily close to 1/2 if the Hessian matrix of the cost function at the minimizing point has all its eigenvalues to the right of 1/2, the cost function is sufficiently smooth, and a sufficiently high-order approximation of the derivative is used.**

***Index Terms*—L-mixing processes, limit theorem for moments, linear stochastic systems, maximal inequalities, recursive estimation.**

I. Introduction

THE aim of this paper is to prove a rate of convergence theorem for a class of stochastic approximation processes for function minimization developed by Spall in [25]. The main feature of Spall's method is a new way of estimating the gradient using only two measurements at properly selected random parameter values. One of the main application areas of simultaneous perturbation stochastic approximation (SPSA) is direct stochastic adaptive control (cf., [25]).

A Kiefer–Wolfowitz method using randomized differences was first proposed in [17, Sec. III-B]. However, the number of measurements required for this method is the same as for the standard Kiefer–Wolfowitz method, i.e., twice the number of dimension. The argument behind the use of randomized differences was that by their use, sensitivity with respect to bias is reduced. A random direction Kiefer–Wolfowitz method with just two measurements has been proposed in [19, Sec. II-C.5], using random unit vectors as perturbing vectors; see also [20].

The main advances of this paper are that a crucial boundedness hypothesis, given as [25, Assumptions A3 and A5, p. 335] is removed by forcing the estimator to stay in a bounded domain, and we get the rate of convergence of higher order moments of the estimation error. Finally, in Section IV, higher order SPSA methods are developed and their convergence properties are established.

The boundedness of the estimator sequence is *a priori* assumed also in [17], and a similar but weaker boundedness condition is assumed in [19]. Namely, it is assumed that the estimator sequence visits a certain fixed domain infinitely often (cf., Theorem 2.3.5).

The rate that we get is—not surprisingly—identical to what appears in the CLT (cf., [25, Proposition 2]). It is expected that the present rate of convergence result will play a role in solving practical problems, such as the analysis of the effect of parametric uncertainty on performance. The present paper also extends results given in [11] in which a complete analysis for the convergence of Ljung's scheme (cf., [5], [4], [21], [1]) with resetting is given.

Higher order Kiefer–Wolfowitz methods were first considered in [7], where the rate of mean-squared error for globally defined procedures had been established. The results of Section IV complement these results: we consider SPSA methods rather than Kiefer–Wolfowitz methods, the procedures are localized using a resetting mechanism, and the rate of higher order moments of the error process is established.

The analysis given in [25] is based on the early work of Fabian in connection with the Kiefer–Wolfowitz method (cf., [6]). It is hoped that the ideas that had emerged since then in the theory of recursive identification yield a more transparent proof that can be adapted to future needs.

The present paper also complements recent results of [2], where an almost sure convergence rate has been given [2, Th. 3] for a modified version of the SPSA algorithm.

II. The Problem Formulation

The p-dimensional Euclidean space will be denoted by $\mathbb{R}^p$. The Euclidean norm of a vector x will be denoted by $|x|$. The operator norm of a matrix A will be denoted be $||A||$, i.e., $||A|| = \sup_{x\neq 0} |Ax|/|x|$. Finally a convention: in the various estimations below we shall frequently have constants which depend only on the constants that appear in the conditions below. These constants will be called system constants.

We consider the following problem: minimize the function $L(\theta)$ defined for $\theta \in D$, where $D \subset \mathbb{R}^p$ is an open domain, for which only noise corrupted measurements are available, given in the form

$$L(\theta) + \epsilon_n$$

Manuscript received November 1, 1996; revised September 2, 1997 and July 9, 1998. Recommended by Associate Editor, G. G. Yin.

The author is with the Computer and Automation Institute, Hungarian Academy of Sciences, H-1111 Budapest, Kende 13-17, Hungary.

Publisher Item Identifier S 0018-9286(99)03941-0.

where $\epsilon_n = \epsilon_n(\omega)$ is a random variable over some probability space $(\Omega, \mathcal{F}, \mathcal{P})$. It is assumed that the measured values of $L(\theta) + \epsilon_n$ can be obtained for each n, ω via a physical experiment, and if necessary the experiment can be repeated. Note that the measurement-noise ϵ_n does not depend on θ, thus we have what is called a state-independent noise. The extension of the results of this paper to state-dependent noise is possible, but a number of additional technical details have to be clarified. This is the subject of a forthcoming paper.

Condition 2.1: The function $L(\theta)$ is three times continuously differentiable with respect to θ for $\theta \in D$. Let $K < \infty$ denote an upper bound of the operator norms of the derivatives up to order three. It is assumed that the minimizing value of $L(\theta)$ is unique in D and will be denoted by θ^*.

Condition 2.2: The measurement noise process $(\epsilon_n) = (\epsilon_n(\omega))$ is assumed to be a zero-mean L-mixing process with respect to a pair of families of σ-fields $(\mathcal{F}_n, \mathcal{F}_n^+)$.

For the definition of L-mixing, cf., Definition 5.2 of the Appendix and [9]. Actually we need less than L-mixing, namely it is sufficient that an "improved Hölder inequality" given below in (1), (cf., [9, Lemma 2.3] for continuous time), is satisfied. Analogous inequalities for uniformly mixing stationary sequences are given in [16] and for strong mixing stationary sequences in [3]. The quoted result is the following: 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

$$|Ex_t y| \leq 2\gamma_q(t-s, x) E^{1/r}|y|^r \tag{1}$$

for every $1 \leq q, r \leq \infty$ such that $1/q + 1/r = 1$.

The class of L-mixing processes has been first systematically studied in [9] and later in [13]. The usefulness of L-mixing processes in stochastic systems theory has been demonstrated in a number of papers, a survey of which is given in [10] and [12]. A basic example of L-mixing processes is obtained as follows: let $(e_n), n \geq 0$ be an M-bounded (cf., Definition 5.1 of the Appendix) independent sequence of real- or vector-valued random variables and define a vector-valued process (y_n) by

$$\begin{aligned} x_{n+1} &= Ax_n + Be_n \\ y_n &= Cx_n \end{aligned} \tag{2}$$

with A stable and $x_0 = 0$. Then it is easy to see that the process $(y_n), n \geq 0$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ defined as

$$\mathcal{F}_n = \sigma\{e_i \colon i \leq n\}, \qquad \mathcal{F}_n^+ = \sigma\{e_i \colon i > n\}. \tag{3}$$

An important property of L-mixing processes is that if (x_n) is a vector-valued L-mixing process and $F(x)$ is a function such that the function itself and the function $G(x,y) = (F(x) - F(y))/|x - y|, x \neq y$ are polynomially increasing, then the process $F(x_n)$ is also L-mixing. Thus the sum or product of L-mixing processes is L-mixing. Furthermore, if (e_n) is L-mixing and (y_n) is generated by (2), then (y_n) is also L-mixing.

To minimize $L(\theta)$ we need an estimator of its gradient, denoted by

$$G(\theta) = L_\theta(\theta). \tag{4}$$

The conventional finite difference approximation of partial derivatives that requires a large number of function evaluations is replaced by an approximation using simultaneous random perturbations of the components of θ. Let k denote the iteration time for the stochastic gradient algorithm to be developed. At time k we take a random vector over some probability space $(\Omega', \mathcal{F}', \mathcal{P}')$

$$\Delta_k(\omega') = (\Delta_{k1}, \cdots, \Delta_{kp})^T.$$

Condition 2.3: $\Delta_{ki} = \Delta_{ki}(\omega')$ is a double sequence of i.i.d., symmetrically distributed, bounded random variables such that for any $m \geq 1$ $E_{P'}\Delta_{ki}^{-m} < \infty$ (cf., [25, Sec. III]).

Remark: Note that the processes $(\epsilon_n) = (\epsilon_n(\omega))$ and $(\Delta_n) = (\Delta_n(\omega'))$ are defined on different probability spaces. From now on we consider the product space $(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', P \times P')$ and write $\epsilon_n(\omega, \omega') = \epsilon_n(\omega)$ and $\Delta_n(\omega, \omega') = \Delta_n(\omega)$. Thus the processes (ϵ_n) and (Δ_n) are independent over $(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mathcal{P} \times \mathcal{P}')$. Mathematical expectation will be always meant to be taken over the mentioned probability space unless otherwise stated.

A standard perturbation that will be used in the rest of the paper is the double sequence Δ_{ki}

$$P'(\Delta_{ki}(\omega') = +1) = 1/2 \quad P'(\Delta_{ki}(\omega') = -1) = 1/2.$$

Let

$$\mathcal{F}_n' = \sigma\{\Delta_k, k = 1, \cdots, n\}$$

and

$$\mathcal{F}_n'^+ = \sigma\{\Delta_k, k = n+1, n+2, \cdots\}. \tag{5}$$

Since Δ_{ki} is an i.i.d. sequence of bounded random variables it follows that (Δ_n) is L-mixing with respect to $(\mathcal{F}_n', \mathcal{F}_n'^+)$.

Now let $0 < c_k \leq 1$ be a fixed sequence of positive numbers and let D_0 be a compact convex domain specified in Condition 2.4 below. For each $\theta \in D_0$ we take two measurements that are denoted by $M_k^+(\theta) = M_k^+(\theta, \omega, \omega')$ and $M_k^-(\theta) = M_k^-(\theta, \omega, \omega')$, defined as

$$\begin{aligned} M_k^+(\theta) &= L(\theta + c_k\Delta_k) + \epsilon_{2k-1} \\ M_k^-(\theta) &= L(\theta - c_k\Delta_k) + \epsilon_{2k}. \end{aligned}$$

Then the estimator of the gradient at time k for $\theta \in D_0$ is defined as

$$\begin{aligned} H(k,\theta) &= H(k,\theta,\omega,\omega') \\ &= \left[\frac{M_k^+(\theta) - M_k^-(\theta)}{2c_k\Delta_{k1}}, \cdots, \frac{M_k^+(\theta) - M_k^-(\theta)}{2c_k\Delta_{kp}}\right]^T. \end{aligned} \tag{6}$$

A convenient representation is obtained if we define the random vector

$$\Delta_k^{-1} = \Delta_k^{-1}(\omega') = (\Delta_{k1}^{-1}, \cdots, \Delta_{kp}^{-1})^T.$$

Then the gradient estimator at $\theta \in D_0$ can be written as

$$H(k,\theta) = (M_k^+(\theta) - M_k^-(\theta))(2c_k)^{-1}\Delta_k^{-1}. \tag{7}$$

The common numerator of these differences can be written as $(L(\theta + c_k\Delta_k)) - L(\theta - c_k\Delta_k)) + \epsilon_{1k}$, where ϵ_{1k} is the compound measurement error defined by

$$\epsilon_{1k} = \epsilon_{2k-1} - \epsilon_{2k}. \tag{8}$$

Define $\overline{\mathcal{F}}_k = \mathcal{F}_{2k}$ $\overline{\mathcal{F}}_k^+ = \mathcal{F}_{2k}^+$. Then it is easy to see that (ϵ_{1k}) is L-mixing with respect to $(\overline{\mathcal{F}}_k, \overline{\mathcal{F}}_k^+)$.

Thus we can write $H(k,\theta)$ as

$$\begin{aligned}(L(\theta + c_k\Delta_k) - L(\theta - c_k\Delta_k))(2c_k)^{-1}\Delta_k^{-1} \\ + \epsilon_{1k}(2c_k)^{-1}\Delta_k^{-1}.\end{aligned} \tag{9}$$

A standard choice for c_k is

$$c_k = c/k^\gamma \quad \text{with some} \quad \gamma > 0. \tag{10}$$

Note that in [25] the condition imposed on the measurement noise is expressed in terms of the compound measurement noise ϵ_{1k}, which is assumed to be a martingale difference process (cf., the condition preceding [25, (2.2), p. 333]). The condition of the present paper given as Condition 2.2 is a possible alternative to Spall's condition.

The ordinary differential equation (ODE)

$$\dot{y}_t = -\frac{a}{t}G(y_t) \quad y_s = \xi, \qquad a > 0 \tag{11}$$

$t \geq s$ will be called the associated differential equation. $G(y)$ is defined in D and it has continuous partial derivatives up to second order. Under the condition above (11) has a unique solution in $[s,\infty)$ which we denote by $y(t,s,\xi)$. It is well known that $y(t,s,\xi)$ is a continuously differentiable function of (t,s,ξ).

Condition 2.4: Let $D_0 \subset \operatorname{int} D$ denote a compact convex domain such that $\theta^* \in \operatorname{int} D_0$, and the closure of the neighborhood of D_0 of radius $c > 0$, denoted as $D(c)$, is inside D. For every $\xi \in D_0, t > s \geq 1$ $y(t,s,\xi) \in D$ is defined and we have with some $C_0, \alpha > 0$

$$\left|\left|\frac{\partial}{\partial\xi}y(t,s,\xi)\right|\right| \leq C_0(s/t)^{a\alpha}. \tag{12}$$

Here $||\cdot||$ denotes the operator norm of a matrix. Furthermore, we assume that the initial condition $\xi \epsilon \operatorname{int} D_{00} \subset \operatorname{int} D_0$, where D_{00} is a compact domain which is invariant for (11) and that for any $t > s \geq 1$

$$y(t,s,D_{00}) = \{y(t,s,x)\colon x \in D_{00}\} \subset \operatorname{int} D_{00}$$

and for $\xi \in D_{00}$ the solution trajectory $y(t,s,\xi)$ converges to θ^*.

Inequality (12) is equivalent to the condition that the differential equation

$$\frac{d}{dv}z_v = -aG(z_v), \qquad z_u = \xi \tag{13}$$

is exponentially asymptotically stable with exponent α, i.e., if the solution of (13) is denoted by $z(v,u,\xi)$, then we have

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

This is obtained by a simple change of time-scale $t = e^v s = e^u$. It is easy to see that in Condition 2.4 α can be taken to be the smallest eigenvalue of the Hessian-matrix of L at $\theta = \theta^*$.

The SPSA Method: Let a_k be a fixed sequence of positive numbers with a_k denoting the stepsize at time k. Then, in the original form of Spall's method, we start with an initial estimate $\hat{\theta}_0$ and then a sequence of estimated parameters, denoted by $\hat{\theta}_{k+1} = \hat{\theta}_{k+1}(\omega,\omega'), k = 0,1,\cdots$, is generated recursively by

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_{k+1}H(k+1,\hat{\theta}_k). \tag{14}$$

A standard choice is $a_k = a/k^\delta$ with $0 < \delta \leq 1$, and $a > 0$.

The central limit theorem (CLT) given as [25, Proposition 2] indicates that for any fixed choice of γ the best rate of convergence is obtained if we choose $\delta = 1$. Thus, in the sequel we shall assume that $a_k = a/k$. There seems to be no technical difficulty to extend the analysis presented in this paper to the case $0 < \delta < 1$.

The almost sure convergence of the estimator process for the case when the noise is a martingale difference process has been established in [25] using results of [22]. In the same paper asymptotic normality of a properly scaled estimation error process is established by a nontrivial application of [6]. The scaling is nonstandard compared to classical statistical theory: assuming $a_k = a/k$ and $c_k = c/k^\gamma$ a normal limit distribution of $k^{(1-2\gamma)/2}(\hat{\theta}_k - \theta^*)$ exists for $1/6 < \gamma < 1/2$.

A main *advance* of the present paper is that the "boundedness conditions" [25, Conditions A3 and A5] repeatedly criticized in the literature on recursive identification (cf., [1, remarks, pp. 46 and 47]) are removed by the application of a resetting or truncation mechanism as described below. A further advance is that by the application of the methods of [11] we get an upper bound for the rate of convergence of the moments of the error process, which is likely to be tight (cf., the remarks following the theorem).

Resetting: Assume that the initial estimate $\hat{\theta}_0$ is in D_{00}. Define, following (14) a tentative value $\hat{\theta}_{k+1}^+ = \hat{\theta}_k - a_{k+1}H(k+1,\hat{\theta}_k)$, and then set

$$\hat{\theta}_{k+1} = \hat{\theta}_{k+1}^+ \quad \text{when} \quad \hat{\theta}_{k+1}^+ \in D_0$$

and

$$\hat{\theta}_{k+1} = \hat{\theta}_0 \quad \text{when} \quad \hat{\theta}_{k+1}^+ \notin D_0. \tag{15}$$

Remark: The above resetting mechanism seems to lose information obtained up to time k. However, if the noise sequence is "untypical" so that it drives the estimator out of the domain D_0, then we can not expect to extract much information. An alternative resetting mechanism would be to chose $\hat{\theta}_{k+1} = \hat{\theta}_k$. However, $\hat{\theta}_k$ may be at a position from which the solution of the ODE does not converge to θ^* at all, or hits the boundary of the truncation domain, and this would force the estimator to be reset infinitely many times.

The choice of the initial estimate $\hat{\theta}_0$ and of the set D_0 requires some *a priori* knowledge of the problem. However, if the associated ODE is globally asymptotically stable in D, the domain of definition of L, then for any initial condition $\xi = \hat{\theta}_0 \in D$ a "sufficiently large" D_0 is a suitable truncation domain. This approach is practical if the parameterization of

the domain D is simple and it is easy to describe what is a large domain D_0. As an example consider the problem of system identification. Let the parameter θ denote the system's parameters of a stable, linear single-input/single-output (SISO) system. If we use the balanced parameterization of Ober (cf., [24]), then the parameter-space is very simple, and the proposed procedure is feasible.

In the theorem below the notation $O_M(\cdot)$ means that the $L_q(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mathcal{P} \times \mathcal{P}')$-norm of the left-hand side decreases with the rate given on the right hand side for any $q \geq 1$.

Theorem 2.1: Let $\beta = \min(4\gamma, 1-2\gamma) > 0$. Assume that the smallest eigenvalue of the Hessian matrix of L at $\theta = \theta^*$, denoted by α, satisfies $a\alpha > \beta/2$. Then under the conditions above, i.e., Conditions 2.1–2.4 we have

$$\hat{\theta}_k - \theta^* = O_M(k^{-\beta/2}). \tag{16}$$

For $a\alpha < \beta/2$ we have $\hat{\theta}_k - \theta^* = O_M(k^{-a\alpha})$. Finally for $a\alpha = \beta/2$ we have for any $\epsilon > 0$ $\hat{\theta}_k - \theta^* = O_M(k^{-\beta/2+\epsilon})$.

The value of $\beta = \min(4\gamma, 1-2\gamma)$ is maximized for $4\gamma = 1-2\gamma$, from which we get $\gamma = 1/6$ and $\beta = 2/3$. The best rate is then $k^{-1/3}$.

Remark: The proof of the theorem yields the following stronger result: let $1 < q < \infty$ and define

$$\tilde{\theta}^*_{q^n} = \sup_{q^n \leq k < q^{n+1}} |\hat{\theta}_k - \theta^*|$$

then, for $a\alpha > \beta/2$, $\tilde{\theta}^*_{q^n} = O_M(q^{-n\beta/2})$.

Remark: The role of the relation between $a\alpha$ and β can be roughly explained as follows: it will be proved in Lemma 3.2 that the estimator sequence locally tracks the solution of the ODE given by (11) with an error $O_M(k^{-\beta/2})$, irrespective of the value of α. But the solution of the ODE converges to θ^* with a rate $O(t^{-a\alpha})$. If this rate is better than the rate of the local error, then the latter will dominate. Otherwise, it is the other way around.

Remark: The theorem does not claim that the given convergence rate is sharp. But the CLT given as [25, Proposition 2] indicates that in the case $a\alpha > \beta/2$, the convergence rate given in the theorem is sharp, indeed.

Remark: Note that we do not need to have the lower bound $1/6 < \gamma$ as in [25, Proposition 2], this is only needed for the asymptotic normality result of [25]. The reason for this is that for $1/6 > \gamma$ the contribution of the third-order term of the Taylor-series expansion (called $J^{\Delta 3}(r, \theta)$ in (27) below) dominates the error process $\overline{J}_r$ [cf., (30) below], hence the stochastic effect is suppressed and the existence of a limiting distribution is not ensured.

For the next theorem we consider an alternative noise condition, in which *no dependence* structure is imposed on (ϵ_n). This theorem is based on an observation in [2].

Condition 2.5: The measurement noise process $(\epsilon_n) = (\epsilon_n(\omega))$ is assumed to be a bounded sequence of random variables.

Theorem 2.2: Let the conditions of Theorem 2.1 be satisfied so that Condition 2.2 is replaced by Condition 2.5. Then the conclusions of Theorem 2.1 remain valid.

The validity of the conclusion of Theorem 2.1 under such weak condition imposed on the noise is a surprising result. This remarkable feature of SPSA has been first observed in [2] in the context of almost sure convergence. Note that no similar result is known for the standard Kiefer–Wolfowitz method. However randomized Kiefer–Wolfowitz methods (cf., [17] and [19, Sec. II-C.5]) do exhibit similar robustness with respect to noise.

An interesting special case is when there is no measurement noise. Then we have a standard optimization problem which is solved by a randomization method. The use of SPSA is justified for large scale problems with low precision requirements.

Theorem 2.3: Let $\epsilon_n = 0$ for all n. Choose $\gamma \geq 1/2$ and assume that the smallest eigenvalue of the Hessian matrix of L at $\theta = \theta^*$, denoted by α, satisfies $a\alpha > 1/2$. Then under Conditions 2.1, 2.3, and 2.4 we have

$$\hat{\theta}_k - \theta^* = O_M(k^{-1/2}). \tag{17}$$

III. The Analysis

Step 1—Continuous-Time Embedding: Some of the calculations to follow are easier to carry out in continuous time, hence we embed our discrete-time data and procedure into a continuous-time data and procedure. Consider the piecewise linear curve $\hat{\theta}^{c+}_t$ defined for $k \leq t \leq k+1$ as $\hat{\theta}^{c+}_t = (t-k)\hat{\theta}^+_{k+1} + (k+1-t)\hat{\theta}_k$, where $\hat{\theta}^+_{k+1}$ is the tentative value of the estimator computed by (14). If $\hat{\theta}^+_{k+1} \in D_0$, then the straight line $\hat{\theta}^{c+}_t$ will lie completely in D_0, since D_0 is convex, and then we set $\hat{\theta}^c_t = \hat{\theta}^{c+}_t$. On the other hand if $\hat{\theta}^+_{k+1} \notin D_0$, then let $\tau = \tau(k) < k+1$ denote the moment when $\hat{\theta}^{c+}_t$ first hits the boundary of D_0. Let us then reset $\hat{\theta}^+_\tau$ to $\hat{\theta}_0$, i.e., we define $\hat{\theta}^c_t = \hat{\theta}^{c+}_t$ for $k \leq t < \tau$ and $\hat{\theta}^c_\tau = \hat{\theta}_0$. In the period $\tau \leq t \leq k+1$ we keep $\hat{\theta}^c_t$ constant: $\hat{\theta}^c_t = \hat{\theta}_0$.

Furthermore, let the piecewise constant continuous-time extension of $(H(k+1, \theta))$, denoted as $(H(t, \theta)) = (H^c(t, \theta)), t > 0$, be defined as $H^c(t, \theta) = H(k+1, \theta)$ for $k < t \leq k+1$. We define the continuous-time extensions of the sequences $(a_k), (c_k), (\Delta_k), \epsilon_{1k}$ similarly.

Set $\mathcal{F}^c_t = \mathcal{F}_{n+1}, \mathcal{F}^{c+}_t = \mathcal{F}^+_{n+1}$ for $n < t \leq n+1$. Then it is easy to see, using, e.g., inequality (74) of the Appendix, that the process (ϵ_{1t}) is L-mixing with respect to $(\mathcal{F}^c_t, \mathcal{F}^{c+}_t)$. Similarly, defining $\mathcal{F}'^c_t = \mathcal{F}'_{n+1}, \mathcal{F}'^{c+}_t = \mathcal{F}'^+_{n+1}$ for $n < t \leq n+1$, the process (Δ_t) is L-mixing with respect to $(\mathcal{F}_t{}'^c, \mathcal{F}'^{c+}_t)$.

Lemma 3.1: In the period $k \leq t < \tau(k)$ the straight line $\hat{\theta}^c_t$ will be the solution of a differential equation of the form

$$\frac{d}{dt}\hat{\theta}^c_t = -\frac{a}{t}(H^c(t, \hat{\theta}^c_t) + \delta H^c(t)) \tag{18}$$

where $\delta H^c(t) = O_M(k^{-1+\gamma})$. If the measurement noise is zero, then we have $\delta H^c(t) = O_M(k^{-1})$ for any γ.

Proof: The correction term $\delta H^c(t)$ is defined by the requirement that the right-hand side yields the constant velocity $\hat{\theta}^+_{k+1} - \hat{\theta}_k$, i.e., we require $-(a/t)(H^c(t, \hat{\theta}^c_t) + \delta H^c(t)) = -(a/(k+1))H(k+1, \hat{\theta}_k)$. From here we get

$$\delta H^c(t) = \frac{t}{k+1} H(k+1, \hat{\theta}_k) - H^c(t, \hat{\theta}^c_t).$$

The time-dependent random field $H(k,\theta)$ is easily seen to be Lipschitz-continuous with respect to θ. First note that for $\theta \neq \theta'$ we have

$$L(\theta + c_k\Delta_k) - L(\theta' + c_k\Delta_k) = \int_0^1 L_\theta(\overline{\theta}^+(\lambda))(\theta - \theta')\, d\lambda$$

where $\overline{\theta}^+(\lambda) = \lambda(\theta + c_k\Delta_k) + (1-\lambda)(\theta' + c_k\Delta_k)$. A similar expression can be obtained for $L(\theta - c_k\Delta_k) - L(\theta' - c_k\Delta_k)$. Since the measurement noise is independent of θ we get, subtracting the latter expression from the former one

$$H(k,\theta) - H(k,\theta') = \int_0^1 (L_\theta(\overline{\theta}^+(\lambda)) - L_\theta(\overline{\theta}^-(\lambda))) \cdot (2c_k)^{-1}\Delta_k^{-1}\, d\lambda(\theta - \theta').$$

Using a second-order Taylor-series expansion for $L_\theta(\overline{\theta}^+(\lambda) - L_\theta(\overline{\theta}^-(\lambda))$ around $\lambda\theta + (1-\lambda)\theta'$ it is easy to see that the Euclidean norm of the integrand is bounded by pK, and thus the Lipschitz continuity of H is established with a Lipschitz constant pK.

To get an upper bound for $\delta H^c(t)$ we subtract and add $-(t/(k+1))H(k+1,\hat{\theta}_t^c) = -(t/(k+1))H^c(t,\hat{\theta}_t^c)$ to get

$$|\delta H^c(t)| \le \frac{t}{k+1} pK|\hat{\theta}_k - \hat{\theta}_t^c| + \left|\left(\frac{t}{k+1} - 1\right) H^c(t,\hat{\theta}_t^c)\right|. \tag{19}$$

For the first term we have

$$|\hat{\theta}_k - \hat{\theta}_t^c| \le |\hat{\theta}_k - \hat{\theta}_{k+1}^+| = \left|\frac{a}{k+1} H(k+1,\theta_k)\right| \le \sup_{\theta\in D_0}\left|\frac{a}{k+1} H(k+1,\theta)\right|.$$

Now consider the expression for the random field $(H(k,\theta))$ given in (9). We have

$$|(L(\theta + c_k\Delta_k) - L(\theta - c_k\Delta_k))(2c_k)^{-1}\Delta_k^{-1}| \le pK$$

with some deterministic constant K, uniformly in θ for $\theta \in D_0$. On the other hand

$$\epsilon_{1k}(2c_k)^{-1}\Delta_k^{-1} = O_M(c_k^{-1})$$

thus we get altogether that for the standard choice $c_k = c/k^\gamma$

$$H^*(k+1) = \sup_{\theta\in D_0} |H(k+1,\theta)| = O_M(c_{k+1}^{-1}) = O_M(k^\gamma).$$

We conclude that the contribution of the first term in (19) is $O_M(k^{-1+\gamma})$.

For the second term in (19) we get $H^c(t,\hat{\theta}_t^c) = H(k+1,\hat{\theta}_t^c)$ and the norm of the latter is majorated by $H^*(k+1)$. Since $((t/k+1) - 1) = O(k^{-1})$ we conclude that the second term in (19) is also of the order of magnitude $O_M(k^{-1+\gamma})$, and thus finally we get that

$$\delta H^c(t) = O_M(k^{-1+\gamma}) \tag{20}$$

which completes the proof.

If the measurement noise is zero, then $H^*(k+1) = O_M(1)$, and thus we get an error term $O_M(k^{-1})$ for any γ.

Step 2—Analysis on Finite Intervals: Let $\sigma \ge 1$, and $q > 1$. We consider the trajectory $\hat{\theta}_t^c$ on the interval $[\sigma, q\sigma \wedge \tau(\sigma))$. Let $\tau(\sigma)$ denote the first moment after σ, at which $\hat{\theta}_t^c$ hits the boundary of D_0. If $\hat{\theta}_t^c$ does not hit the boundary of D_0 at all, then we set $\tau(\sigma) = \infty$. Further, let $\overline{y}_t$ denote the solution of (11) with initial condition $\overline{y}_\sigma = \hat{\theta}_\sigma = \theta$, i.e., $\overline{y}_t = y(t,\sigma,\theta)$. A main technical tool used in the proof is the development of a locally uniform upper bound for the increments $|\hat{\theta}_t^c - \overline{y}_t|$. Note that the validity of the lemma is independent of α, the smallest eigenvalue of the Hessian-matrix of L at $\theta = \theta^*$. Let

$$I_\sigma(q) = \sup_{\sigma\le t\le q\sigma\wedge\tau(\sigma)} |\hat{\theta}_t^c - \overline{y}_t|$$
$$I_s^*(q) = \sup_{s\le\sigma\le qs} I_\sigma(q).$$

Lemma 3.2: Let $\delta = 1$, $0 < \gamma < 1/2$ and let $\beta = \min(4\gamma, 1-2\gamma)$. Then we have $I_s^*(q) = O_M(s^{-\beta/2})$. If the measurement noise is zero, then taking $\gamma \ge 1/2$ we get $I_s^*(q) = O_M(s^{-1/2})$.

Proof: First fix σ. Since $\hat{\theta}_\sigma^c = \overline{y}_\sigma = \theta$ we have for $\sigma \le t \le q\sigma \wedge \tau(\sigma)$

$$\begin{aligned}\hat{\theta}_t^c - \overline{y}_t &= \int_\sigma^t \frac{a}{r}(-H^c(r,\hat{\theta}_r^c) - \delta H^c(r) + G(\overline{y}_r))\, dr \\ &= \int_\sigma^t \frac{a}{r}((-H^c(r,\hat{\theta}_r^c) - \delta H^c(r) + G(\hat{\theta}_r^c)) \\ &\quad - (G(\hat{\theta}_r^c) - G(\overline{y}_r)))\, dr \\ &= \int_\sigma^t \frac{a}{r}(J_r - (G(\hat{\theta}_r^c) - G(\overline{y}_r)))\, dr \end{aligned} \tag{21}$$

where J_r is the cumulative error defined as

$$J_r = -H^c(r,\hat{\theta}_r^c) - \delta H^c(r) + G(\hat{\theta}_r^c). \tag{22}$$

This error process will be decomposed according to the source of the error. First write

$$J_r^{\Delta\epsilon} = -H^c(r,\hat{\theta}_r^c) + G(\hat{\theta}_r^c). \tag{23}$$

To analyze $J_r^{\Delta\epsilon}$ substitute the free variable θ for $\hat{\theta}_r^c$. Define

$$J^{\Delta\epsilon}(r,\theta) = -H^c(r,\theta) + G(\theta).$$

Then $J_r^{\Delta\epsilon} = J^{\Delta\epsilon}(r,\hat{\theta}_r^c)$. Now we continue to decompose $J^{\Delta\epsilon}(r,\hat{\theta}_r^c)$ taking into account the representation of $H^c(r,\theta) = H(k+1,\theta)$ given in (7): we write

$$\begin{aligned} J^{\Delta\epsilon}(r,\theta) &= J^\Delta(r,\theta) + J_r^\epsilon \\ J^\Delta(r,\theta) &= -(L(\theta + c_r\Delta_r) - L(\theta - c_r\Delta_r))(2c_r)^{-1} \\ &\quad \cdot \Delta_r^{-1} + G(\theta) \\ J_r^\epsilon &= -\epsilon_{1r}(2c_r)^{-1}\Delta_r^{-1}. \end{aligned} \tag{24}$$

The first term, $J^\Delta(r,\theta)$, will be further decomposed using a third-order Taylor-series expansion. Write

$$\begin{aligned}&(L(\theta + c_r\Delta_r) - L(\theta))\cdot(2c_r)^{-1}\Delta_r^{-1} \\ &\quad = L_\theta(\theta)^T(c_r\Delta_r)(2c_r)^{-1}\Delta_r^{-1} + \tfrac{1}{2}(c_r\Delta_r)^T \\ &\qquad \cdot L_{\theta\theta}(\theta)(c_r\Delta_r)\cdot(2c_r)^{-1}\Delta_r^{-1} + \tfrac{1}{6}\int_0^1 L_{\theta\theta\theta}(\overline{\theta}^+(\lambda)) \\ &\qquad \cdot d\lambda * (c_r\Delta_r) * (c_r\Delta_r) * (c_r\Delta_r)\cdot(2c_r)^{-1}\Delta_r^{-1}\end{aligned}$$

where $\overline{\theta}^{+}(\lambda)=\lambda(\theta+c_r\Delta_r)+(1-\lambda)\theta$ and $*$ denotes appropriate tensor products. A similar expansion can be obtained for $L(\theta-c_r\Delta_r)-L(\theta)$. Subtracting the latter expansion from the one given above the first- and third-order term will be doubled and the second-order terms will be cancelled, thus we get

$$\begin{aligned}-&(L(\theta+c_r\Delta_r)-L(\theta-c_r\Delta_r))\cdot(2c_r)^{-1}\Delta_r^{-1}+G(\theta)\\ &=(-(L_\theta(\theta)^T(2c_r\Delta_r)(2c_r)^{-1}\Delta_r^{-1}-G(\theta))\\ &\quad-\tfrac{1}{6}\int_0^1+L_{\theta\theta\theta}(\overline{\theta}^{+}((\lambda))\,d\lambda*(c_r\Delta_r)*(c_r\Delta_r)*(c_r\Delta_r)\\ &\quad\cdot(2c_r)^{-1}\Delta_r^{-1}. \end{aligned}\tag{25}$$

Note that we have

$$\begin{aligned}&L_\theta(\theta)^T(2c_r\Delta_r)\cdot(2c_r)^{-1}\Delta_r^{-1}\\ &\quad=(L_\theta(\theta)^T\Delta_r)\cdot\Delta_r^{-1}=\Delta_r^{-1}(\Delta_r^TL_\theta(\theta))\\ &\quad=\Delta_r^{-1}\Delta_r^TG(\theta)\end{aligned}$$

and thus the first term on the right-hand side of (25) can be written as

$$J^{\Delta1}(r,\theta)=-(\Delta_r^{-1}\Delta_r^T-I)G(\theta).\tag{26}$$

In $J^{\Delta1}(r,\theta)=J^{\Delta1}(r,\theta,\omega')$ randomness is purely due to the random perturbation of the parameters. The conditions imposed on Δ_r imply that

$$E_{I'}(-\Delta_r^{-1}\Delta_r^T+I)=0.$$

Since Δ_{ki} is L-mixing with respect to $(\mathcal{F}'_n,\mathcal{F}'^{+}_n)$, we get that $\Delta_r^{-1}\Delta_r^T$ is also L-mixing with respect to $(\mathcal{F}'_n,\mathcal{F}'^{+}_n)$, and we conclude that $J^{\Delta1}(r,\theta)$ is a zero-mean L-mixing process with respect to $(\mathcal{F}'_n,\mathcal{F}'^{+}_n)$. The same holds trivially for its first three derivatives with respect to θ.

For the third-order term we introduce the notations

$$\begin{aligned}J^{\Delta3}(r,\theta)=&-\tfrac{1}{6}\int_0^1(L_{\theta\theta\theta}(\overline{\theta}^{+}(\lambda))+L_{\theta\theta\theta}(\overline{\theta}^{-}(\lambda))\,d\lambda(c_r\Delta_r)\\ &*(c_r\Delta_r)*(c_r\Delta_r)\cdot(2c_r)^{-1}\Delta_r^{-1}.\end{aligned}\tag{27}$$

It is easy to see that when θ is restricted to $D_0(c)$ we have

$$\sup_{\theta\in D_0(c)}|J^{\Delta3}(r,\theta)|=O(c_r^2)=O(k^{-2\gamma})$$

and hence

$$J_r^{\Delta3}=J^{\Delta3}(r,\hat{\theta}_r^c)=O(c_r^2)=O(k^{-2\gamma}).\tag{28}$$

Thus we have arrived at the following decomposition of the error process $J_r=J_r$:

$$J_r=J^{\Delta_1}(r,\hat{\theta}_r^c)+J_r^{\Delta3}+J_r^{\epsilon}+J_r^{c}.\tag{29}$$

For the sake of further reference the relevant terms are summarized as follows:

$$\begin{aligned}J^{\Delta1}(r,\theta)&=-(\Delta_r^{-1}\Delta_r^T-I)G(\theta)\\ J_r^{\Delta3}&=-J^{\Delta3}(r,\hat{\theta}_r^c)\\ J^{\Delta3}(r,\theta)&=-\tfrac{1}{6}\int_0^1(L_{\theta\theta\theta}(\hat{\theta}^{+}(\lambda)+L_{\theta\theta\theta}(\hat{\theta}^{-}(\lambda)\\ &\quad\cdot d\lambda*(c_r\Delta_r)*(c_r\Delta_r)*(c_r\Delta_r)\\ &\quad\cdot(2c_r)^{-1}\Delta_r^{-1}\\ J_r^{\epsilon}&=-\epsilon_{1r}(2c_r)^{-1}\Delta_r^{-1}\\ J_r^{c}&=-\delta H^c(r).\end{aligned}$$

We further decompose $J^{\Delta1}(r,\hat{\theta}_r^c)$ in order to simplify the randomness present in $\hat{\theta}_r^c$. Let us write

$$J^{\Delta1}(r,\hat{\theta}_r^c)=J^{\Delta1}(r,\overline{y}_r)+(J^{\Delta1}(r,\hat{\theta}_r^c)-J^{\Delta1}(r,\overline{y}_r)).$$

Since θ is restricted to a compact domain in which the first and second derivatives of L are bounded, and the components of Δ_r have absolute value equal to one, it follows from the definition of $J^{\Delta1}(r,\theta)$ [cf., (3.9)] that $J^{\Delta1}(r,\theta)$ is Lipschitz-continuous, say $|J^{\Delta1}(r,\hat{\theta}_r^c)-J^{\Delta1}(r,\overline{y}_r)|\le K'|\hat{\theta}_r^c-\overline{y}_r|$, where $K'<\infty$ is a deterministic constant, depending only on K and p. Now write

$$J_r^{\Delta1}=J^{\Delta1}(r,\overline{y}_r)=J^{\Delta1}(r,y(r,\sigma,\hat{\theta}_\sigma^c)).$$

The advantage of this approximation is that the randomness of $\overline{y}_r$ is purely due to the randomness of the initial condition $\hat{\theta}_\sigma^c$. Define the modified error process

$$\overline{J}_r=J_r^{\Delta_1}+J_r^{\Delta3}+J_r^{\epsilon}+J_r^{c}.\tag{30}$$

Substituting into (21) and taking into account the inequality $|G(\hat{\theta}_r^c)-G(\overline{y}_r)|\le K|\hat{\theta}_r-\overline{y}_r|$, we get that $|\hat{\theta}_t^c-\overline{y}_t|$ is majorated by

$$\begin{aligned}&\left|\int_\sigma^t\frac{a}{r}(J_r^{\Delta1}+J_r^{\Delta3}+J_r^{\epsilon}+J_r^{c})dr\right|\\ &\qquad+\int_\sigma^t\frac{a}{r}(K'+K)|\hat{\theta}_r^c-\overline{y}_r|\,dr.\end{aligned}\tag{31}$$

Now we are in a position to apply the Bellman–Gronwall lemma with σ fixed. But first we need to get a tight upper bound for the first term on the right-hand side.

Step 3—Estimation of the Constant Term: Let us consider the expressions

$$\begin{aligned}\delta_s^*(J^{\Delta1})&=\sup_{\substack{\sigma\le t\le q\sigma\wedge\tau(\sigma)\\ \theta\in D_0\\ s\le\sigma\le qs}}\left|\int_\sigma^t\frac{a}{r}J^{\Delta1}(r,y(r,\sigma,\theta))\,dr\right|\\ \delta_s^*(J^{\Delta3})&=\sup_{\substack{\sigma\le t\le q\sigma\wedge\tau(\sigma)|\\ s\le\sigma\le qs}}\left|\int_\sigma^t\frac{a}{r}J^{\Delta3}(r,\hat{\theta}_r^c)\,dr\right|\\ \delta_s^*(J^{\epsilon})&=\sup_{\substack{\sigma\le t\le q\sigma\wedge\tau(\sigma)\\ s\le\sigma\le qs}}\left|\int_\sigma^t\frac{a}{r}J_r^{\epsilon}\,dr\,dr\right|\\ \delta_s^*(J^{c})&=\sup_{\substack{\sigma\le t\le q\sigma\wedge\tau(\sigma)\\ s\le\sigma\le qs}}\left|\int_\sigma^t\frac{a}{r}J_r^{c}\,dr\right|.\end{aligned}$$

Note that since $\theta\in D_0$ Condition 2.4 above implies that $y(r,\sigma,\theta)$ is defined for all $r\ge\sigma$ and hence the definition of $\delta_s^*(J^{\Delta1})$ is correct. Defining the compound error term

$$\delta_s^*=\delta_s^*(J^{\Delta1})+\delta_s^*(J^{\Delta3})+\delta_s^*(J^{\epsilon})+\delta^*(J^{c})\tag{32}$$

we get from (31) for any $\sigma\le t\le q\sigma\wedge\tau(\sigma)$ with $s\le\sigma\le qs$

$$|\hat{\theta}_t^c-\overline{y}_t|\le\delta_s^*+\int_\sigma^t\frac{1}{r}(K'+K)|\hat{\theta}_r^c-\overline{y}_r|\,dr.\tag{33}$$

Estimation of $\delta_s^*(J^{\Delta 1})$: It is easy to see that for fixed σ and θ the process

$$J^{\Delta 1}(r, y(r,\sigma,\theta)) = -(\Delta_r^{-1}(\omega')\Delta_r^T(\omega') - I)G(y(r,\sigma,\theta))$$

$r \geq \sigma$ and its first two partial derivatives with respect to θ are zero-mean L-mixing processes over $(\Omega', \mathcal{F}', \mathcal{P}')$. Using this fact, [11, Lemma 3.1] implies that

$$\delta_s^*(J^{\Delta 1}) = O_M(s^{-1/2}) \tag{34}$$

in the sense that the $L_q(\Omega', \mathcal{F}', \mathcal{P}')$-norm of the left-hand side decreases with the rate given on the right-hand side for any $q \geq 1$. Obviously, if we consider $\delta_s^*(J^{\Delta 1})$ as a random variable over the product space $(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mathcal{P} \times \mathcal{P}')$, the same proposition holds.

Estimation of $\delta_s^*(J^{\Delta 3})$: Using (28) and the assumption $c_r = c/(k+1)^\gamma \leq c/r^\gamma$ we get

$$\begin{aligned}\delta_s^*(J^{\Delta 3}) &\leq \left| \int_s^{q^2 s} \frac{a}{r} \cdot C c_r^2 \, dr \right| \leq \left| \int_s^{q^2 s} aC \frac{c^2}{r^{1+2\gamma}} \, dr \right| \\ &\leq C_1 s^{-2\gamma}\end{aligned} \tag{35}$$

with some $C_1 > 0$.

Estimation of $\delta_s^*(J^\epsilon)$: We have [cf., (24)]

$$\delta_s^*(J^\epsilon) = \sup_{\substack{\sigma \leq t \leq q\sigma \wedge \tau(\sigma) \\ s \leq \sigma \leq qs}} \left| \int_s^t \frac{a}{r} \epsilon_r (2c_r)^{-1} \Delta_r^{-1} \, dr \right|.$$

By Condition 2.2 the process (ϵ_{1r}) considered as a process over $(\Omega, \mathcal{F}, \mathcal{P})$ is a zero-mean L-mixing process with respect to $(\mathcal{F}_r, \mathcal{F}_r^+)$. It is easy conclude that (ϵ_{1r}) considered as a process over the product space $(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mathcal{P} \times \mathcal{P}')$, is a zero-mean L-mixing process with respect to $(\mathcal{F}_r \times \mathcal{F}_r', \mathcal{F}_r^+ \times \mathcal{F}_r'^+)$. Similarly (Δ_r^{-1}), considered as a process over the product space $(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mathcal{P} \times \mathcal{P}')$, is a zero-mean L-mixing process with respect to $(\mathcal{F}_r \times \mathcal{F}_r', \mathcal{F}_r^+ \times \mathcal{F}_r'^+)$. Since ϵ_{1r} and Δ_r^{-1} are independent we conclude that $(\epsilon_{1r}\Delta_r^{-1})$ is a zero-mean L-mixing process over $(\Omega \times \Omega', \mathcal{F} \times \mathcal{F}', \mathcal{P} \times \mathcal{P}')$ with respect to $(\mathcal{F}_r \times \mathcal{F}_r', \mathcal{F}_r^+ \times \mathcal{F}_r'^+)$.

Thus for fixed σ we can estimate the moments of the integral on the right-hand side using the maximal inequality given as [9, Th. 5.1] and restated as Theorem 5.1 in the Appendix. We get for any $m \geq 1$ and fixed σ

$$\begin{aligned}&E^{1/2m} \sup_{\sigma \leq t \leq q\sigma} \left| \int_\sigma^t \frac{a}{r} \epsilon_{1r} (2c_r)^{-1} \Delta_r^{-1} \, dr \right|^{2m} \\ &\quad \leq C \left(\int_\sigma^{q\sigma} \frac{a^2}{r^2} (2c_r)^{-2} \, dr \right)^{1/2}\end{aligned}$$

where C depends on m and the processes $(\epsilon_{1r}), (\Delta_r)$, but is independent of σ and q when the latter is confined to a bounded interval. Substituting $c_r = c/r^\gamma$ we get for the right-hand side

$$\left(\int_\sigma^{q\sigma} \frac{a^2}{r^2} \frac{r^{2\gamma}}{c^2} \, dr \right)^{1/2} \leq C \sigma^{-1/2+\gamma}.$$

To proceed we need to develop an extension of [11, Lemma 3.1]. Defining

$$\delta_\sigma(J^\epsilon) = \sup_{\sigma \leq t \leq q\sigma} \left| \int_\sigma^t \frac{a}{r} \epsilon_{1r} (2c_r)^{-1} \Delta_r^{-1} \, dr \right| \tag{36}$$

we can write the above inequality as

$$\delta_\sigma(J^\epsilon) = O_M(\sigma^{-1/2+\gamma}). \tag{37}$$

The additional difficulty in estimating $\delta_s^*(J^\epsilon)$ is in the handling of the supremum with respect to σ over a set of dilating intervals. This will be done with the application of an appropriate change of time scale. Let us set $\sigma = e^v$ and define

$$\rho_v(J^\epsilon) = e^{(1/2-\gamma)v} \delta_{e^v}(J^\epsilon). \tag{38}$$

With this notation we have the following proposition.

Lemma 3.3: The processes $\rho_v(J^\epsilon)$ and $|\rho_{v+k}(J^\epsilon) - \rho_v(J^\epsilon)|/|k|^{1/2}, k \neq 0$ are M-bounded.

Proof: We have already shown that $\delta_\sigma := \delta_\sigma(J^\epsilon) = O_M(\sigma^{-1/2+\gamma})$, therefore $\rho_v := \rho_v(J^\epsilon) = O_M(1)$. Let us now take a small $k > 0$ and estimate the moments of $\rho_{v+k} - \rho_v$. Write $e^k = 1 + h$, then we can write

$$\rho_{v+k} - \rho_v = (\sigma(1+h))^{1/2-\gamma} \delta_{\sigma(1+h)} - \sigma^{1/2-\gamma} \delta_\sigma.$$

The difference $\delta_{\sigma(1+h)} - \delta_\sigma$ can obviously be majorated by the sum of the following two terms:

$$\Delta_1 = \sup_{\sigma \leq t \leq \sigma(1+h)} \left| \int_\sigma^t \frac{a}{r} \epsilon_{1r} (2c_r)^{-1} \Delta_r^{-1} \, dr \right| \tag{39}$$

$$\Delta_2 = \sup_{q\sigma \leq t \leq q\sigma(1+h)} \left| \int_{q\sigma}^t \frac{a}{r} \epsilon_{1r} (2c_r)^{-1} \Delta_r^{-1} \, dr \right|. \tag{40}$$

We estimate Δ_1 similarly to $\delta_\sigma(J^\epsilon)$: for any $m \geq 1$

$$\begin{aligned}&E^{1/2m} \sup_{\sigma \leq t \leq \sigma(1+h)} \left| \int_\sigma^t \frac{a}{r} \epsilon_{1r} (2c_r)^{-1} \Delta_r^{-1} \, dr \right|^{2m} \\ &\quad \leq C \left(\int_\sigma^{\sigma(1+h)} \frac{a^2}{r^2} (2c_r)^{-2} \, dr \right)^{1/2}\end{aligned}$$

where C depends on m and the processes $(\epsilon_{1r}), (\Delta_r)$, but is independent of σ and q when the latter is confined to a bounded interval. Substituting $c_r = c/r^\gamma$ we get for the right-hand side

$$\begin{aligned}&\left(\int_\sigma^{\sigma(1+h)} \frac{a^2}{r^2} \frac{r^{2\gamma}}{c^2} \, dr \right)^{1/2} \\ &\quad = \left(\frac{a^2}{c^2} \frac{1}{2\gamma - 1} \cdot r^{2\gamma-1} |_\sigma^{\sigma(1+h)} \right)^{1/2} \\ &\quad = C'(\sigma^{2\gamma-1}(1 - (1+h)^{2\gamma-1}))^{1/2}\end{aligned}$$

where C' depends only on c and γ. Thus we get

$$\Delta_1 = O_M(\sigma^{-1/2+\gamma} h^{1/2}). \tag{41}$$

Similar estimates can be obtained for Δ_2, thus we finally get

$$\delta_{\sigma(1+h)} - \delta_\sigma = O_M(\sigma^{-1/2+\gamma} h^{1/2}). \tag{42}$$

Now write

$$\begin{aligned}\rho_{v+k} - \rho_v &= (\sigma(1+h))^{1/2-\gamma} \delta_{\sigma(1+h)} - \sigma^{1/2-\gamma} \delta_\sigma \\ &= (\sigma(1+h))^{1/2-\gamma} (\delta_{\sigma(1+h)} - \delta_\sigma) \\ &\quad + ((\sigma(1+h))^{1/2-\gamma} - \sigma^{1/2-\gamma}) \delta_\sigma.\end{aligned}$$

Then, by (42) the first term on the right-hand side is $O_M(h^{1/2})$, when h is confined to a bounded interval. For the

second term we get by (37) and by the elementary inequality $((\sigma(1+h))^{1/2-\gamma} - \sigma^{1/2-\gamma}) = O(\sigma^{1/2-\gamma}h)$ that its order of magnitude is $O_M(h)$. Since for small k we have $h = O(k)$, we finally get

$$\rho_{v+k} - \rho_v = O_M(k^{1/2}). \tag{43}$$

Thus Lemma 3.3 has been proved.

Now let $q>1$ be fixed and let $h = q-1$. Set $e^k = 1+h$ and $s = e^w$. Then applying the maximal inequality given as Theorem 5.2 of the Appendix to the congruent compact intervals $[w, w+k]$ with varying w, we get

$$\rho_w^* = \sup_{w \le v \le w+k} \rho_v = O_M(1). \tag{44}$$

Observing that $\delta_{e^v}(J^\epsilon) = e^{(-1/2+\gamma)v}\rho_v(J^\epsilon)$ [cf., (38)], we immediately get that $\delta_s^*(J^\epsilon) = \delta_{e^w}^*(J^\epsilon) \le e^{(-1/2+\gamma)w}\rho_w^*(J^\epsilon) = O_M(e^{(-1/2+\gamma)w}) = O_M(s^{-1/2+\gamma})$ thus we conclude that

$$\delta_s^*(J^\epsilon) = O_M(s^{-1/2+\gamma}). \tag{45}$$

If the measurement noise is zero, then obviously $\delta_s^*(J^\epsilon) = 0$ for any γ.

Estimation of $\delta_s^*(J^c)$: Using (20) we get

$$\begin{aligned}\delta_s^*(J^c) &\le \sup_{\substack{\sigma \le t \le q\sigma \wedge \tau(\sigma) \\ s \le \sigma \le qs}} \int_\sigma^t \left| \frac{a}{r}\delta H^c(r) \right| dr \\ &\le \int_s^{q^2 s} \frac{a}{r} C r^{-1+\gamma}\, dr \le C s^{-1+\gamma}. \end{aligned} \tag{46}$$

If the measurement noise is zero, then $\delta_s^*(J^c) = O_M(s^{-1})$ for any γ.

Summarizing (34), (35), (45), (46) it is easy to see that for $\gamma < 1/2$ the dominant terms are $\delta_s^*(J^{\Delta 3})$, arising from the approximation error in the numerical differentiation scheme, and $\delta_s^*(J^\epsilon)$, arising from the measurement error. Thus for the compound error term δ_s^* defined under (32) we get

$$\delta_s^* = O(s^{-2\gamma}) + O_M(s^{-1/2+\gamma}). \tag{47}$$

It follows that we can write

$$\delta_s^* = O_M(s^{-\beta/2}) \quad \text{with} \quad \beta = \min(4\gamma, 1-2\gamma). \tag{48}$$

If the measurement noise is zero, then summarizing (34), (35), (45), (46) we see that for $\gamma \ge 1/2$ the dominant term is $\delta_s^*(J^{\Delta 1})$. It follows that we have

$$\delta_s^* = O_M(s^{-1/2}). \tag{49}$$

Now from (33) we get using the Bellman–Gronwall lemma with fixed σ

$$I_\sigma(q) = \sup_{\sigma \le t \le q\sigma \wedge \tau(\sigma)} |\hat{\theta}_t^c - \overline{y}_t| \le \kappa \delta_s^* \tag{50}$$

with

$$\kappa = \exp \int_\sigma^{q\sigma} \frac{1}{r}(K' + K)\, dr = q^{(K'+K)}.$$

Since the right-hand side of (50) is independent of σ, we can take supremum over σ on the left-hand side, and thus Lemma 3.2 is proved.

Step 4—Hitting Probabilities: Let C_s denote the event that a resetting takes place in the interval $[s, qs)$. It has been shown in [11, Lemma 2.3] that Lemma 3.2 implies that $P(C_s) = O(s^{-m})$ for all $m>0$. It follows that in the whole interval $[s, qs)$ we have

$$\sup_{s \le t \le qs} |\hat{\theta}_t^c - \overline{y}_t| \le \kappa \delta_s^* + \chi_{C_s} \cdot K = O_M(s^{-\beta/2}) \tag{51}$$

where K is the diameter of D_0. Note that the validity of this estimate is independent of α.

Step 5—Pasting Together the Interval Estimates: Let us now take a subdivision of $[1, \infty]$ by the points $s_n = q^n$ with some $q>1$. Then

$$\delta_{s_n}^* = O_M(q^{-n\beta/2}). \tag{52}$$

Assume that $a\alpha > \beta/2$. Following the arguments on [11, p. 1208] and using the stability condition imposed on the associated ODE, it follows that with y_t denoting the solution of the associated ODE starting from $\hat{\theta}_0$ we have

$$\sup_{q^n \le t < q^{n+1}} |\hat{\theta}_t^c - y_t| = O_M(q^{-n\beta/2}).$$

Since $\sup_{q^n \le t < q^{n+1}} |y_t - \theta^*| = O(q^{-na\alpha})$, the proposition of the theorem follows. The case $a\alpha < \beta/2$ is handled analogously (cf., [11]), using a corrected version of [11, Lemma 7.4], given as Lemma 5.1 in the Appendix of this paper. The case $a\alpha = \beta/2$ trivially follows from the previous results. The proof of Theorem 2.3 is completed analogously.

Proof of Theorem 2.2: We have to reconsider only the arguments given under the heading "*Estimation of* $\delta_s^*(J^\epsilon)$." The main idea is that the boundedness of $(\epsilon_{1r}(\omega))$ implies that for any fixed ω the process $(\epsilon_{1r}(\omega)\Delta_r^{-1}(\omega'))$ is a zero-mean L-mixing process over $(\Omega', \mathcal{F}', \mathcal{P}')$ with respect to $(\mathcal{F}_r', \mathcal{F}_r'^+)$, uniformly in ω (cf., [9]). Writing

$$\delta_s^*(J^\epsilon) = \delta_s^*(J^\epsilon(\omega, \omega')) \tag{53}$$

we get by the arguments given above that for all $m>1$ we have

$$E_{P'}^{1/2m} |\delta_s^*(J^\epsilon(\omega, \omega'))|^{2m} \le C s^{-1/2+\gamma} \tag{54}$$

where C is independent of ω. Integrating this inequality over Ω with respect to (ω, dP) we get the desired inequality (45), and the rest of the proof is the same.

IV. Higher Order Schemes

The idea behind SPSA can be generalized to get higher order approximations of the gradient. Higher order Kiefer–Wolfowitz methods were first considered in [7], where a special ad hoc approximation scheme is used. In contrast to this we rely on approximation schemes the numerical properties of which are known to be very good. The rate of mean-squared error established in [7] is identical to the rate that we get for higher order moments of the error process. However, better numerical procedures may improve the asymptotic covariance of the estimation error.

Following [8, Ch. 2], let $f(x)$ be a real-valued function of the real variable x, and let $h>0$. Then define the shift operator S_h by

$$(S_h f)(x) = f(x+h) \tag{55}$$

and the central difference operator δ_h by

$$(\delta_h f)(x) = f(x+h/2) - f(x-h/2). \tag{56}$$

Clearly we can write

$$\delta_h = S_h^{1/2} - S_h^{-1/2}. \tag{57}$$

On the other hand, for f analytic, a Taylor-expansion around x gives

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \cdots$$

which can be expressed in the form $f(x+h) = e^{hD}f(x)$, where D is the differentiation operator. But $f(x+h) = (S_h f)(x)$, thus we can write formally

$$S_h = e^{hD}. \tag{58}$$

Combining this with (57), we conclude that

$$\delta_h = e^{hD/2} - e^{-hD/2} = 2\sinh(\tfrac{1}{2}hD) \tag{59}$$

from which we get for D the representation

$$hD = 2\ \sinh^{-1}\ \tfrac{1}{2}\ \delta_h. \tag{60}$$

The first few terms of the Taylor-series expansion of the right-hand side gives the approximation

$$hD = 2\left\{\left(\frac{\delta_h}{2}\right) - \frac{1}{2}\left(\frac{\delta_h}{2}\right)^3\frac{1}{3} + \frac{1}{2}\frac{3}{4}\left(\frac{\delta_h}{2}\right)^5\frac{1}{5}\cdots\right\}. \tag{61}$$

Let the $2m$th-order formal Taylor-series expansion of the right-hand side of (60) be denoted by $P_{2m}(\delta_h)$. It is stated (cf., [8, p. 22]) that for a function f having $2m+1$ continuous derivatives, we have

$$hf'(x) = P_{2m}(\delta_h)f(x) + \frac{h^{2m+1}(-1)^m(m!)^2 f^{(2m+1)}(\xi)}{(2m+1)!} \tag{62}$$

where in the last term ξ is in the range $x-mh \le \xi \le x+mh$.

This approximation of the derivative is favored in numerical analysis because of its good convergence properties, namely the coefficients of the expansion of the right-hand side of (60) decay faster to zero, than for other expansions.

For a multivariable function $L(\theta)$ we approximate the partial derivatives analogously. For this we fix $h>0$ and define the central difference operator $\delta_{v,h}$ in the direction v by applying the operator δ_h to the function $f(x) = L(\theta+xv)$ of the scalar variable x. Thus if the function L is $2m+1$ continuously differentiable, then we have

$$hf'(0) = hv^T\frac{\partial L}{\partial \theta}(\theta) = P_{2m}(\delta_{v,h})L(\theta) + O(h^{2m+1}) \tag{63}$$

when θ is restricted to a compact domain.

Now take $v=\Delta_k$ random as in Section II and let $0<c_k\le 1$ be a fixed sequence of positive numbers. The gradient estimator will be based on measurements taken for $i = 1,\cdots,m$ at $\theta \pm (i-1/2)c_k\Delta_k$. A single measurement has the form $L(\theta\pm(i-1/2)c_k\Delta_k) + \epsilon_k^{\pm(i-1/2)}$, where $\epsilon_k^{\pm(i-1/2)}$ is the measurement error. Note that for $m=1$ the positions at which measurements are taken are halfway between θ and the positions taken by standard SPSA. We assume that for $i=1,\cdots,m$ we have $\theta\pm(i-1/2)c_k \in D_0$. Let $M(\theta)$ denote a generic measurement taken at θ, i.e., $M(\theta) = L(\theta)+\epsilon$.

Then the gradient estimator at $\theta \in D_0$ is defined as

$$H(k,\theta) = P_{2m}(\delta_{\Delta_k,c_k})M(\theta)c_k^{-1}\Delta_k^{-1} \tag{64}$$

where the notation $P_{2m}(\delta_{\Delta_k,c_k})M(\theta)$ is self-explanatory. We define the estimator sequence as in (14) replacing $H(k,\theta)$ by the expression given above in (64) and using the same resetting rule.

The analysis of higher order SPSA methods is analogous to that of the second-order SPSA method given above. In view of the assumed independence of $\epsilon_k^{\pm i}$ and Δ_k we have

$$EP_{2m}(\delta_{\Delta_k,c_k})M(\theta)c_k^{-1}\Delta_k^{-1} = EP_{2m}(\delta_{\Delta_k,c_k})L(\theta)c_k^{-1}\Delta_k^{-1}.$$

Furthermore by (63) the latter expression is equal to

$$E\left(c_k\Delta_k^T\frac{\partial L}{\partial\theta}(\theta) + O(c_k^{2m+1})\right)c_k^{-1}\Delta_k^{-1} = \frac{\partial L}{\partial\theta}(\theta) + O(c_k^{2m+1}).$$

The effect of using higher order approximation schemes is that the residual term $J^{\Delta 3}(r,\theta)$ defined under (27) will be replaced by a higher order residual term $J^{\Delta,2m+1}(r,\theta)$, for which we have [cf., (28)]

$$J_r^{\Delta,2m+1} = O(c_r^{2m}) = O(k^{-2m\gamma}). \tag{65}$$

The estimation of $\delta_s^*(J^{\Delta,2m+1})$ will proceed as in (35): we get

$$\delta_s^*(J^{\Delta,2m+1}) \le \left|\int_s^{q^2 s}\frac{a}{r}\cdot Cc_r^{2m}\,dr\right| \le \left|\int_s^{q^2 s} aC\,\frac{c^{2m}}{r^{1+2m\gamma}}\,dr\right| \le C_1 s^{-2m\gamma} \tag{66}$$

with some $C_1>0$. Summarizing (34), (66), (45), (46) it is again easy to see that for $\gamma<1/2$ the dominant terms are $\delta_s^*(J^{\Delta,2m+1})$ and $\delta_s^*(J^\epsilon)$, the latter arising from the measurement error. Thus for the compound error term δ_s^* [cf., (32)] which is now defined as

$$\delta_s^* = \delta_s^*(J^{\Delta 1}) + \delta_s^*(J^{\Delta,2m+1}) + \delta_s^*(J^\epsilon) + \delta^*(J^c) \tag{67}$$

we get the estimation

$$\delta_s^* = O(s^{-2m\gamma}) + O_M(s^{-1/2+\gamma}). \tag{68}$$

It follows that we can write

$$\delta_s^* = O_M(s^{-\beta/2}) \quad \text{with} \quad \beta = \min(4m\gamma, 1-2\gamma). \tag{69}$$

Thus we get in general the following result.

Theorem 4.1: Let $\beta = \min(4m\gamma, 1-2\gamma) > 0$. Assume that the smallest eigenvalue of the Hessian matrix of L at $\theta = \theta^*$, denoted by α, satisfies $a\alpha > \beta/2$. Assume that the conditions of Theorem 2.1, i.e., Conditions 2.1–2.4, are satisfied with the following additions: the function L is $2m+1$ times continuously differentiable in D, and in Condition 2.4 the neighborhood of D_0 of radius $(m+1)c/2$ is inside D. Then for the estimator sequence $\hat{\theta}_k$ defined by (14), combined with a resetting mechanism, with H being defined under (64), we have

$$\hat{\theta}_k - \theta^* = O_M(k^{-\beta/2}). \tag{70}$$

For $a\alpha < \beta/2$ we have $\hat{\theta}_k - \theta^* = O_M(k^{-a\alpha})$. Finally for $a\alpha = \beta/2$ we have for any $\epsilon > 0$ $\hat{\theta}_k - \theta^* = O_M(k^{-a\alpha+\epsilon})$.

The value of $\beta = \min(4m\gamma, 1-2\gamma)$ is maximized for $4m\gamma = 1 - 2\gamma$, from which we get $\gamma = 1/(4m+2)$ and $\beta = 4m/(4m+2) = 2m/(2m+1)$. The best rate that is obtained is

$$\hat{\theta}_k - \theta^* = O_M(k^{-m/(2m+1)}).$$

This rate can be arbitrarily close to $1/2$ if m is sufficiently large.

APPENDIX

In this section, we summarize some definitions and earlier results that have been used in the paper. Let a probability space $(\Omega, \mathcal{F}, P)$ be given and let $D \subset \mathbb{R}^p$ be an open domain. A parameter-dependent stochastic process $(x_n(\theta)), n \geq 0$, or equivalently a time-varying random field, is a sequence of measurable mappings for $n \geq 0$ from $(\Omega \times D, \mathcal{F} \times \mathcal{B}(D))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Here $\mathcal{B}(D)$ denotes the σ-field of Borel sets of D.

Definition 5.1: We say that the $\mathbb{R}^m$-valued parameter-dependent stochastic process $(x_n(\theta))$ is M-bounded if for all $1 \leq q < \infty$

$$M_q(x) := \sup_{\substack{n \geq 0 \\ \theta \in D}} E^{1/q}|x_n(\theta)|^q < \infty.$$

If (x_n) is M-bounded we shall also write $x_n = O_M(1)$. Similarly, if c_n is a positive sequence we write $x_n = O_M(c_n)$ if $x_n/c_n = O_M(1)$. The definition trivially extends to parameter-independent processes.

The first part of the following result was stated in [11, Lemma 7.4]. The second part of the quoted lemma was not correctly stated and is therefore restated and proved here.

Lemma 5.1: Let $(u_n), n \geq 0$ be an M-bounded process and define a process (x_n) by

$$x_{n+1} = \lambda x_n + \rho^n u_n, \qquad x_0 = 0 \tag{71}$$

where $0 < \lambda < \rho < 1$. Then for any $m \geq 1$ we have

$$E^{1/m}|x_n|^m \leq \frac{\rho^n}{\rho - \lambda} M_m(u).$$

On the other hand if $0 < \rho < \lambda < 1$, then we have

$$E^{1/m}|x_n|^m \leq \frac{\lambda^n}{\lambda - \rho} M_m(u).$$

Proof: Let $0 < \lambda < \rho < 1$, and set $z_n = \rho^{-n} x_n$. Then we have after multiplying (71) by $\rho^{-(n+1)}$

$$z_{n+1} = \lambda \rho^{-1} z_n + \rho^{-1} u_n$$

which can be solved explicitly for z_n

$$z_n = \sum_{i=0}^{n-1} (\lambda \rho^{-1})^{n-1-i} \rho^{-1} u_i.$$

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 derivation is as follows: write

$$x_n = \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} u_{n-1-i}.$$

Then we have

$$\begin{aligned} E^{1/m}|x_n|^m &\leq \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} E^{1/m}|u_{n-1-i}|^m \\ &\leq \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} M_m(u). \end{aligned} \tag{72}$$

Thus it is sufficient to establish that for $0 < \lambda < \rho < 1$

$$\sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} \leq \frac{\rho^n}{\rho - \lambda}$$

and this has been done above. 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 the same estimate for $E^{1/m}|x_n|^m$, but the role of λ and ρ is interchanged thus we get

$$E^{1/m}|x_n|^m \leq \frac{\lambda^n}{\lambda - \rho} M_m(u).$$

Let $(\mathcal{F}_n), n \geq 0$ be a monotone increasing family of σ-algebras, and $(\mathcal{F}_n^+), n \geq 0$ be a monotone decreasing family of σ-algebras. We assume that for all $n \geq 0, \mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent. A standard example is

$$\mathcal{F}_n = \sigma\{e_i \colon i \leq n\} \quad \mathcal{F}_n^+ = \sigma\{e_i \colon i > n\} \tag{73}$$

where $(e_i), i \geq 0$ is an independent sequence of random variables.

Definition 5.2: An $\mathbb{R}^m$-valued stochastic process $(x_n), n \geq 0$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ if it is $\mathcal{F}_n$-adapted, M-bounded, and with τ being a nonnegative integer and

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

we have for any $1 \leq q < \infty$

$$\Gamma_q(x) = \sum_{\tau=0}^{\infty} \gamma_q(\tau, x) < \infty.$$

Note that in the definition of $\Gamma_q(x)$ the moment $\tau = 0$ is included, so that $\Gamma_q(x) \neq 0$ even if the x_n's constitute an independent sequence of random variables not all of which are constants.

The continuous-time extension of Definition 5.1 is straightforward. The extension of the concept of L-mixing for continuous-time processes requires an additional technical condition. Thus let a pair of families of σ-algebras $(\mathcal{F}_t, \mathcal{F}_t^+)$ be given such that: 1) $\mathcal{F}_t \subset \mathcal{F}$ is monotone increasing; 2) $\mathcal{F}_t^+ \subset \mathcal{F}$ is monotone decreasing and $\mathcal{F}_t^+$ is right-continuous in t, i.e., $\mathcal{F}_s^+ = \sigma\{\cup_{0<\epsilon} \mathcal{F}_{s+\epsilon}^+\}$; 3) $\mathcal{F}_t$ and $\mathcal{F}_t^+$ are independent for all t.

Definition 5.3: A stochastic process $(x_t), t \geq 0$ is L-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, if it is $\mathcal{F}_t$-adapted, M-bounded, and with

$$\gamma_q(\tau, x) = \sup_{t \geq \tau} E^{1/q}|x_t - E(x_t|\mathcal{F}_{t-\tau}^+)|^q, \qquad \tau \geq 0$$

we have for any $1 \leq q < \infty$

$$\Gamma_q(x) = \int_0^\infty \gamma_q(\tau, x)\, d\tau < \infty.$$

Although $\gamma_q(\tau, x)$ is in general not monotone decreasing in τ, we have (cf., [9, (2.1)]) for $1 \leq q < \infty, \tau \leq \tau'$

$$\gamma_q(\tau', x) \leq 2\gamma_q(\tau, x). \tag{74}$$

A fundamental technical tool in estimation theory is a moment inequality given as [9, Th. 1.1]. Based on this result, and using a continuous-time extension of a basic inequality due to Móricz (cf., [23]), we get the following maximal inequality given as [9, Th. 5.1].

Theorem 5.1: Let $(x_t), t \geq 0$ be a real-valued L-mixing process with $Ex_t = 0$ for all t and let (f_t) be a deterministic function that is locally in $L_2[0, \infty)$. Then we have for all $1 < m < \infty$

$$E^{1/2m} \sup_{1 \leq T' \leq T} \left| \int_0^{T'} f_t x_t dt \right|^{2m} \leq C'_m \left(\int_0^T f_t^2 dt \right)^{1/2} M_{2m}^{1/2}(x) \Gamma_{2m}^{1/2}(x)$$

where C'_m depends only on m.

The theorem obviously extends to vector-valued processes, weighted by matrix-valued f_t.

An important technical tool is an inequality that provides an upper bound for the maximal value of random fields. To formulate this let $0 < \alpha \leq 1$ and define the time-varying random field $(\Delta x_n / \Delta^\alpha \theta)$ by

$$\Delta x_n / \Delta^\alpha \theta(\theta, \theta + h) = |x_n(\theta + h) - x_n(\theta)|/|h|^\alpha$$

for $n \geq 0, \theta \neq \theta + h \in D$.

Definition 5.4: The random field $(x_n(\theta))$ is M-Hölder-continuous in θ with exponent α, where $0 < \alpha \leq 1$, if the random field $\Delta x / \Delta^\alpha \theta$ is M-bounded, i.e., if for all $1 \leq q < \infty$ we have

$$M_q(\Delta x / \Delta^\alpha \theta) = \sup_{\substack{n \geq 0 \\ \theta \neq \theta + h \in D}} E^{1/q}|x_n(\theta + h) - x_n(\theta)|^q / |h|^\alpha < \infty.$$

For $\alpha = 1$ we say that the random field is M-Lipschitz-continuous in θ.

Let us assume that $(x_n(\theta))$ is measurable, separable, M-bounded, and M-Hölder continuous in θ for $\theta \in D$. By Kolmogorov's theorem (cf., [14]) the realizations of $(x_n(\theta))$ are continuous in θ with probability one. Hence, for $D_0 \subset D$ being a compact domain, we can define for almost all ω

$$x_n^* = \max_{\theta \in D_0} |x_n(\theta)|. \tag{75}$$

The quoted result also gives an upper bound for the expectation of the continuity modulus of $(x_n(\theta))$, which in turn can be used to estimate the moments of x_n^*. An upper bound was already derived in [18, Lemma 7.15, Ch II], and a simple extension of that result yields the following result, given as [9, Th. 3.4]:

Theorem 5.2: Assume that $(x_n(\theta))$ is a measurable, separable, M-bounded random field, which is also M-Hölder-continuous with exponent α for $\theta \in D \subset \mathbb{R}^p$. Let x^* be the random variable defined in (75). Then we have for all $q \geq 1$ and $r > p/\alpha$

$$E^{1/q}(x^*)^q \leq C(M_{qr}(x) + M_{qr}(\Delta x / \Delta^\alpha \theta)) \tag{76}$$

where C depends only on α, p, q, r and D, D^0.

ACKNOWLEDGMENT

The author would like to thank J. C. Spall and J. L. Maryak for initiating and cooperating in this research, the reviewers for providing a number of useful comments, and the editors for bringing attention to reference [17] and to the method of [19, Sec. II-C.5].

REFERENCES

[1] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. Berlin, Germany: Springer-Verlag, 1990.
[2] H. F. Chen, T. E. Duncan, and B. Pasik-Duncan, "A stochastic approximation algorithm with random differences," in *Proc. 13th Triennal IFAC World Congr.*, San Francisco, CA, USA, 1996, pp. 493–496, J. Gertler, J. B. Cruz, and M. Peshkin, Ed.
[3] Yu. A. Davydov, "Convergence of distributions generated by stationary stochastic processes," *Theory Probab. Appl.*, vol. 13, pp. 691–696, 1968.
[4] D. P. Djereveckii and A. L. Fradkov, "Application of the theory of Markov-processes to the analysis of the dynamics of adaptation algorithms," *Automat. Remote Contr.*, vol. 2, pp. 39–48, 1974.
[5] ———, *Applied Theory of Discrete Adaptive Control Systems* (in Russian). Moscow, Russia: Nauka, 1981.
[6] V. Fabian, "On asymptotic normality in stochastic approximation," *Ann. Math. Stat.*, vol. 39, pp. 1327–1332, 1968.
[7] ———, "Stochastic approximation of minima with improved asymptotic speed," *Ann. Math. Stat.*, vol. 38, pp. 191–200, 1968.
[8] L. Fox, *Two-Point Boundary Problems in Ordinary Differential Equations*. Oxford, U.K.: Clarendons, 1957.
[9] L. Gerencsér, "On a class of mixing processes," *Stochastics*, vol. 26, pp. 165–191, 1989.

[10] _____, "Strong approximation results in estimation and adaptive control," in *Topics in Stochastic Systems: Modeling, Estimation and Adaptive Control*, L. Gerensér and P. E. Caines, Ed. Berlin, Germany: Springer-Verlag, 1991, pp. 268–299.
[11] _____, "Rate of convergence of recursive estimators," *SIAM J. Contr. Optim.*, vol. 30, no. 5, pp. 1200–1227, 1992.
[12] L. Gerencsér and J. Rissanen, "Asymptotics of predictive stochastic complexity: From parametric to nonparametric models," in *New Directions in Time-Series Analysis, Part II Proc. 1990 IMA Workshop*, E. Parzen, D. Brillinger, M. Rosenblatt, M. Taqqu, J. Geweke, and P. E. Caines, Ed. Minneapolis, MN: Institute of Mathematics and its Applications, 1993, pp. 93–112.
[13] H. Hjalmarsson, "Aspects on incomplete modeling in system identification," Ph.D. dissertation, Linköping Univ., 1993.
[14] A. Ibragimov and R. Z. Khasminskii, *Statistical Estimation. Asymptotic Theory*. Berlin, Germany: Springer-Verlag, 1981.
[15] I. A. Ibragimov and Yu. A. Linnik, *Independent and Stationary Sequences of Random Variables*. Groningen, The Netherlands: Wolters and Nordhoff, 1971.
[16] J. Koronacki, "Random-seeking methods for the stochastic unconstrained optimization," *Int. J. Contr.*, vol. 21, pp. 517–527, 1975.
[17] N. V. Krylov, *Controlled Diffusion Processes*. Berlin, Germany: Springer-Verlag, 1980.
[18] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Optimization*. New York: Springer, 1978.
[19] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 1997.
[20] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. 22, pp. 551–575, 1977.
[21] M. Métivier and P. Priouret, "Application of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Trans. Inform. Theory*, vol. 30, pp. 140–151, 1984.
[22] F. Móricz, "Moment inequalities and the strong laws of large numbers," *Z. Wahrscheinlichkeitstheorie u. verw. Gebiete*, vol. 35, pp. 299–314, 1974.
[23] R. Ober, "Balanced parametrization of classes of linear systems," *SIAM J. Control ans Optim.*, vol. 29, pp. 1251–1287, 1991.
[24] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 332–341, 1992.
[25] J. C. Spall and J. A. Cristion, "Model-free control of nonlinear stochastic systems with discrete-time measurements," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 1198–1210, 1998.

László Gerencsér received the M.Sc. and doctorate degree in mathematics at the Eötvös Lóránd University (ELTE), Budapest, Hungary, in 1969 and 1970, respectively. In 1976 the candidate of mathematical sciences degree was awarded to him by the Hungarian Academy of Sciences.

Since 1970 he has been with the Computer and Automation Institute of the Hungarian Academy of Sciences, where he is currently heading the Applied Mathematics Laboratory. He held a one-year visiting position at the Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden, in 1986. From 1998 to 1991 he was a Visiting Professor at the Department of Electrical Engineering, McGill University Montreal, Quebec, Canada. He is currently holding a Széchenyi Professorship with the Eötvös Lóránd University, Budapest, Hungary. His main recent research interests include: model-uncertainty and control, stochastic approximation, hidden-Markov models, statistical theory of linear stochastic systems, high accuracy stochastic adaptive control, continuous-time linear stochastic systems, change point detection, and financial mathematics.

Dr. Gerencsér is an Associate Editor for the *SIAM Journal on Control and Optimization.*