Truncated Stochastic Approximation with Moving Bounds: Convergence

Teo Sharia

Department of Mathematics
Royal Holloway, University of London
Egham, Surrey TW20 0EX
e-mail: t.sharia@rhul.ac.uk

Abstract

In this paper we consider a wide class of truncated stochastic approximation procedures. These procedures have three main characteristics: truncations with random moving bounds, a matrix valued random step-size sequence, and a dynamically changing random regression function. We establish convergence and consider several examples to illustrate the results.

Keywords: Stochastic approximation, Recursive estimation, Parameter estimation

1 Introduction

Stochastic approximation (SA) introduced by Robbins and Monro in 1951 ([21]) was created to locate a root of an unknown function when only noisy measurements of the function can be observed. SA quickly became very popular, resulting in interesting new developments and numerous applications across a wide range of disciplines. Comprehensive surveys of the SA technique including some recent developments can be found in [3], [4], [14], [15], [16], [17].

In this paper we consider a wide class of truncated SA procedures with moving random bounds. While we believe that the proposed class of procedures will find its way to a wider range of applications, the main motivation is to accommodate applications to parametric statistical estimation theory. Our class of SA procedures has three main characteristics: truncations with random moving bounds, a matrix-valued random step-size sequence, and a dynamically changing random regression function.

To introduce the main idea, let us first consider the classical problem of finding a unique zero, say z^0, of a real valued function $R(z) : \mathbb{R} \to \mathbb{R}$ when only noisy measurements of R are available. To estimate z^0, consider a sequence defined recursively

as

$$Z_t = \left[\ Z_{t-1} + \gamma_t \left(R(Z_{t-1}) + \varepsilon_t\right)\right]_{\alpha_t}^{\beta_t}, \qquad t = 1, 2, \ldots \tag{1.1}$$

where ε_t is a sequence of zero-mean random variables and γ_t is a deterministic sequence of positive numbers. Here α_t and β_t are random variables with $-\infty \le \alpha_t \le \beta_t \le \infty$ and $[v]_a^b$ is the truncation operator, that is,

$$[v]_a^b = \begin{cases} a & \text{if } v < a, \\ v & \text{if } a \le v \le b, \\ b & \text{if } v > b. \end{cases}$$

We assume that the truncation sequence $[\alpha_t, \beta_t]$ contains z^0 for large values of t. For example, if it is known that z^0 belongs to (α, β), with $-\infty \le \alpha \le \beta \le \infty$, one can consider truncations with expanding bounds to avoid possible singularities at the endpoints of the interval. That is, we can take $[\alpha_t, \beta_t]$ with some sequences $\alpha_t \downarrow \alpha$ and $\beta_t \uparrow \beta$. Truncations with expanding bounds may also be useful to overcome standard restrictions on growth of the corresponding functions.

The most interesting case arises when the truncation interval $[\alpha_t, \beta_t]$ represents our auxiliary knowledge about z^0 at step t, which is incorporated into the procedure through the truncation operator. Consider for example a parametric statistical model. Suppose that $X_1, \ldots, X_t$ are independent and identically distributed random variables and $f(x, \theta)$ is the common probability density function (w.r.t. some σ-finite measure) depending on an unknown parameter $\theta \in \mathbb{R}^m$. Consider the recursive estimation procedure for θ defined by

$$\hat\theta_t = \hat\theta_{t-1} + \frac{1}{t} i(\hat\theta_{t-1})^{-1}\ \frac{f'^T(X_t, \hat\theta_{t-1})}{f(X_t, \hat\theta_{t-1})}, \qquad t \ge 1. \tag{1.2}$$

where f' is the row-vector of partial derivatives of f w.r.t. the components of θ, $i(\theta)$ is the one-step Fisher information matrix, and $\hat\theta_0 \in \mathbb{R}^m$ is some initial value. This estimator was introduced in [23] and studied in [10], [13] and [20]. In particular, it has been shown that under certain conditions the recursive estimator $\hat\theta_t$ is asymptotically equivalent to the maximum likelihood estimator, i.e., it is consistent and asymptotically efficient. The analysis of (1.2) can be conducted by rewriting it in the form of stochastic approximation. Indeed, in the case of (1.2), let us fix θ and let $\gamma_t = 1/t$,

$$R(z) = i(z)^{-1} E^\theta \left\{ \frac{f'^T(X_t, z)}{f(X_t, z)} \right\} \quad \text{and} \quad \varepsilon_t = i(\hat\theta_{t-1})^{-1} \left(\frac{f'^T(X_t, \hat\theta_{t-1})}{f(X_t, \hat\theta_{t-1})} - R(\hat\theta_{t-1}) \right)$$

(E^θ is expectation w.r.t. $f(x, \theta)$). Then, under the usual regularity assumptions, $R(\theta) = 0$ and ε_t is a martingale difference (w.r.t. the filtration $\mathcal{F}_t$ generated by the observations). So, (1.2) is a standard SA of type (1.1) without truncations (i.e., in the one dimensional case, $-\alpha_t = \beta_t = \infty$).

However, the need of truncations may naturally arise from various reasons. One obvious consideration is that the functions in the procedure may only be defined for certain values of the parameter. In this case one would want the procedure to produce points only from this set. Truncations may also be useful when the standard assumptions such as restrictions on the growth rate of the relevant functions are not satisfied. More importantly, truncations may provide a simple tool to achieve an efficient use of information available in the estimation process. This information can be auxiliary information about the parameters, e.g. a set, possibly time dependent, that is known to contain the value of the unknown parameter. Suppose for instance that a consistent (i.e., convergent), but not necessarily efficient auxiliary estimator $\tilde{\theta}_t$ is available having a rate d_t. Then one can consider a truncated procedure with shrinking bounds. The idea is to obtain asymptotically efficient estimator by truncating the recursive procedure in a neighbourhood of θ with $[\alpha_t, \beta_t] = [\tilde{\theta}_t - \delta_t, \tilde{\theta}_t + \delta_t]$, $\delta_t \to 0$. Such a procedure is obviously consistent since $\hat{\theta}_t \in [\tilde{\theta}_t - \delta_t, \tilde{\theta}_t + \delta_t]$ and $\tilde{\theta}_t \pm \delta_t \to \theta$. However, to construct an efficient estimator, care should be taken to ensure that the truncation intervals do not shrink to $\tilde{\theta}_t$ too rapidly, for otherwise $\hat{\theta}_t$ will have the same asymptotic properties as $\tilde{\theta}_t$ (see [29] for details in the case of AR processes). Since this paper is concerned with the convergence, details of this application is not discussed here. However, since the procedures with shrinking bounds are particular cases of the general SA procedure below (see (2.1)), asymptotic distribution and efficiency can be studied in an unified manner using ideas of SA.

Note that the idea of truncations with moving bounds is not new. For example, an idea of truncations with shrinking bounds goes back to [13] and [10]. Truncations with expanding bounds were considered in [1] and also, in the context of recursive parametric estimation, in [24] (see also [29]). Truncations with adaptive truncation sets of the Robbins-Monro SA were introduced in [5], and further explored and extended in [6], [2], [31], [32], [18]. The latter algorithms are designed in such a way, that the procedure is pulled back to a certain pre-specified point or a set, every time the sequence leaves the truncation region. As one can see from (1.1) and (2.1), truncation procedures considered in this paper are quite different from the latter ones and are similar to the the ones introduced in [13], [10] and [1]. A detailed comparison of these two different approaches is given in [1].

Let us now consider a discrete time stochastic processes $X_1, X_2, \ldots$ with the joint distribution depending on an unknown parameter $\theta \in \mathbb{R}^m$. Then one can consider the recursive estimator of θ defined by

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \gamma_t(\hat{\theta}_{t-1})\psi_t(\hat{\theta}_{t-1}), \qquad t \geq 1, \tag{1.3}$$

where $\psi_t(v) = \psi_t(X_1, \ldots, X_t; v)$, $t = 1, 2, \ldots$, are suitably chosen functions which may, in general, depend on the vector of all past and present random variables and have the property that the process $\psi_t(\theta)$ is P^θ- martingale difference, i.e., $E^\theta \{\psi_t(\theta) \mid \mathcal{F}_{t-1}\} = 0$ for each t. For example, if $f_t(x, \theta) = f_t(x, \theta | X_1, \ldots, X_{t-1})$ is

the conditional probability density function of the observation X_t given $X_1, \dots, X_{t-1}$, then one can obtain a likelihood type estimation procedure by choosing $\psi_t(v) = l_t(v) = f_t'(X_t, v)/f_t(X_t, v)$. Asymptotic behaviour of this type of procedures for non i.i.d. models was studied by a number of authors, see e.g., [7], [9], [19], [25] – [28]. Results in [28] show that to obtain an estimator with asymptotically optimal properties, one has to consider a state-dependent matrix-valued random step-size sequence. One possible choice is $\gamma_t(u)$ with the property

$$\gamma_t^{-1}(v) - \gamma_{t-1}^{-1}(v) = E_\theta\{\psi_t(v) l_t^T(v) \mid \mathcal{F}_{t-1}\}$$

In particular, to obtain a recursive procedure which is asymptotically equivalent to the maximum likelihood estimator, one has to consider $l_t(v) = f_t'(X_t, v)/f_t(X_t, v)$ and $\gamma_t(v) = I_t^{-1}(v)$, where $I_t(v)$ is the conditional Fisher information matrix (see [28] for details). To rewrite (1.3) in a SA form, let us assume that θ is an arbitrary but fixed value of the parameter and define

$$R_t(z) = E^\theta \{\psi_t(X_t, z) \mid \mathcal{F}_{t-1}\} \quad \text{and} \quad \varepsilon_t(z) = (\psi_t(X_t, z) - R_t(z)) .$$

Obviously, $R_t(\theta) = 0$ for each t, and $\varepsilon_t(z)$ is a martingale difference.

Therefore, to be able to study these procedures in an unified manner, one needs to consider a SA of the following form

$$Z_t = \left[Z_{t-1} + \gamma_t(Z_{t-1})\big\{R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})\big\}\right]_{U_t}, \quad t = 1, 2, \dots$$

where $R_t(z)$ is predictable with the property that $R_t(z^0) = 0$ for all t's, $\gamma_t(z)$ is a matrix-valued predictable step-size sequence, $U_t \subset \mathbb{R}^m$ is a random sequence of truncation sets, and $Z_0 \in \mathbb{R}^m$ is some starting value (see Section 2 for more details).

To summarise the above, the procedures introduced in this paper have the following features: (1) inhomogeneous random functions R_t; (2) state dependent matrix valued random step sizes; (3) truncations with random and moving (shrinking or expanding) bounds. These are mainly motivated by parametric statistical applications. In particular, (1) is required to include recursive parameter estimation procedures for non i.i.d. models, (2) is needed to guarantee asymptotic optimality and efficiency of statistical estimation, (3) is required to accommodate various different adaptive truncations, including the ones arising by auxiliary estimators. Also, the convergence of these procedures is studied under very general conditions and the results might be of interest even for the procedures without truncations (i.e., when $U_t = \mathbb{R}^m$) and with a deterministic and homogeneous regression function $R_t(z) = R(z)$.

The paper is organised as follows. In Sections 2.2 we prove two theorems on the convergence. The analysis is based on the method of using convergence sets of nonnegative semimartingales. The decomposition into negative and positive parts in these theorems turns out to be very useful in applications (see Example 3 in Section 2.4). In Section 2.3 we give several corollaries in the case of state independent scalar random step-size sequences. In Section 2.4 we consider examples. Proofs of some technical parts are postponed to Section 3.

2 Convergence

2.1 Main objects and notation

Let $(\Omega,\ \mathcal{F}, F = (\mathcal{F}_t)_{t\geq 0},\ P)$ be a stochastic basis satisfying the usual conditions. Suppose that for each $t = 1, 2, \ldots$, we have $(\mathcal{B}(\mathbb{R}^m) \times \mathcal{F})$-measurable functions

$$\begin{aligned} R_t(z) = R_t(z,\omega) &: \mathbb{R}^m \times \Omega \to \mathbb{R}^m \\ \varepsilon_t(z) = \varepsilon_t(z,\omega) &: \mathbb{R}^m \times \Omega \to \mathbb{R}^m \\ \gamma_t(z) = \gamma_t(z,\omega) &: \mathbb{R}^m \times \Omega \to \mathbb{R}^{m\times m} \end{aligned}$$

such that for each $z \in \mathbb{R}^m$, the processes $R_t(z)$ and $\gamma_t(z)$ are predictable, i.e., $R_t(z)$ and $\gamma_t(z)$ are $\mathcal{F}_{t-1}$ measurable for each t. Suppose also that for each $z \in \mathbb{R}^m$, the process $\varepsilon_t(z)$ is a martingale difference, i.e., $\varepsilon_t(z)$ is $\mathcal{F}_t$ measurable and $E\{\varepsilon_t(z) \mid \mathcal{F}_{t-1}\} = 0$. We also assume that

$$R_t(z^0) = 0$$

for each $t = 1, 2, \ldots$, where $z^0 \in \mathbb{R}^m$ is a non-random vector.

Suppose that $h = h(z)$ is a real valued function of $z \in \mathbb{R}^m$. We denote by $h'(z)$ the row-vector of partial derivatives of h with respect to the components of z, that is,

$$h'(z) = \left(\frac{\partial}{\partial z_1}h(z), \ldots, \frac{\partial}{\partial z_m}h(z)\right).$$

Also, we denote by $h''(z)$ the matrix of second partial derivatives. The $m \times m$ identity matrix is denoted by $\mathbf{1}$.

Let $U \subset \mathbb{R}^m$ is a closed convex set and define a truncation operator as a function $\left[z\right]_U : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, such that

$$\left[z\right]_U = \begin{cases} z & \text{if } z \in U \\ z^* & \text{if } z \notin U, \end{cases}$$

where z^* is a point in U, that minimizes the distance to z.

Suppose that $z^0 \in \mathbb{R}^m$. We say that a random sequence of sets $U_t = U_t(\omega)$ $(t = 1, 2, \ldots)$ from $\mathbb{R}^m$ is *admissible* for z^0 if

- for each t and ω, $U_t(\omega)$ is a closed convex subset of $\mathbb{R}^m$;
- for each t and $z \in \mathbb{R}^m$, the truncation $\left[z\right]_{U_t}$ is $\mathcal{F}_t$ measurable;
- $z^0 \in U_t$ eventually, i.e., for almost all ω there exist $t_0(\omega) < \infty$ such that $z^0 \in U_t(\omega)$ whenever $t > t_0(\omega)$.

Assume that $Z_0 \in \mathbb{R}^m$ is some starting value and consider the procedure

$$Z_t = \left[\ Z_{t-1} + \gamma_t(Z_{t-1})\Psi_t(Z_{t-1})\right]_{U_t}, \quad t = 1, 2, \dots \tag{2.1}$$

where U_t is admissible for z^0,

$$\Psi_t(z) = R_t(z) + \varepsilon_t(z),$$

$R_t(z)$, $\varepsilon_t(z)$, $\gamma_t(z)$ are random fields defined above,

$$E\left\{\Psi_t(Z_{t-1}) \mid \mathcal{F}_{t-1}\right\} = R_t(Z_{t-1}), \tag{2.2}$$

$$E\left\{\varepsilon_t^T(Z_{t-1})\varepsilon_t(Z_{t-1}) \mid \mathcal{F}_{t-1}\right\} = \left[E\left\{\varepsilon_t^T(z)\varepsilon_t(z) \mid \mathcal{F}_{t-1}\right\}\right]_{z=Z_{t-1}}, \tag{2.3}$$

and the conditional expectations (2.2) and (2.3) are assumed to be finite.

Remark 2.1 *Note that* (2.2) *in fact means that the sequence* $\varepsilon_t(Z_{t-1})$ *is a martingale difference. Conditions* (2.2) *and* (2.3) *obviously hold if, e.g., the measurement errors* $\varepsilon_t(u)$ *are independent random variables, or if they are state independent. In general, since we assume that all conditional expectations are calculated as integrals w.r.t. corresponding regular conditional probability measures (see the convention below), these conditions can be checked using disintegration formula (see, e.g., Theorem 5.4 in [12]).*

Convention.
• *Everywhere in the present work convergence and all relations between random variables are meant with probability one w.r.t. the measure* P *unless specified otherwise.*
• *A sequence of random variables* $(\zeta_t)_{t\geq 1}$ *has some property* ***eventually*** *if for every* ω *in a set* Ω_0 *of* P *probability 1, the realisation* $\zeta_t(\omega)$ *has this property for all* t *greater than some* $t_0(\omega) < \infty$.
• We assume that all conditional expectations are calculated as integrals w.r.t. corresponding regular conditional probability measures.
• *We will also assume that the* $\inf_{z\in U} h(z)$ *of a real valued function* $h(z)$ *is* 1 *whenever* $U = \emptyset$.

2.2 Convergence theorems

Theorem 2.2 *Let* Z_t *be a process defined by* (2.1), (2.2) *and* (2.3), *with an admissible for* $z^0 \in \mathbb{R}^m$ *truncation sequence* U_t. *Let* $V(u) : \mathbb{R}^m \longrightarrow \mathbb{R}$ *be a real valued nonnegative function having continuous and bounded partial second derivatives. Denote*

$$\triangle_t = Z_t - z^0$$

and suppose that the following conditions are satisfied.

(L)

$$V\Big(\triangle_t\Big) \leq V\Big(\triangle_{t-1} + \gamma_t(Z_{t-1})\Psi_t(Z_{t-1})\Big)$$

eventually.

(S)

$$\sum_{t=1}^{\infty}(1+V(\triangle_{t-1}))^{-1}\left[\mathcal{N}_t(\triangle_{t-1})\right]^+ < \infty, \qquad P\text{-a.s.} \tag{2.4}$$

where

$$\begin{aligned}\mathcal{N}_t(u) &= V'(u)\gamma_t(z^0+u)R_t(z^0+u)\\ &\quad +\frac{1}{2}\sup_v \|V''(v)\| E\left\{\|\gamma_t(z^0+u)\Psi_t(z^0+u)\|^2 \mid \mathcal{F}_{t-1}\right\}.\end{aligned}$$

Then $V(Z_t - z^0)$ converges (P-a.s.) to a finite limit for any initial value Z_0. Furthermore,

$$\sum_{t=1}^{\infty}\left[\mathcal{N}_t(\triangle_{t-1})\right]^- < \infty, \qquad P\text{-a.s.} \tag{2.5}$$

Proof. As always (see the convention in 2.1), convergence and all relations between random variables are meant with probability one w.r.t. the measure P unless specified otherwise.

From condition (**L**), using the Taylor expansion,

$$\begin{aligned}V(\triangle_t) &\leq V(\triangle_{t-1}) + V'(\triangle_{t-1})\gamma_t(z^0+\triangle_{t-1})\Psi_t(z^0+\triangle_{t-1})\\ &+\frac{1}{2}\left[\gamma_t(z^0+\triangle_{t-1})\Psi_t(z^0+\triangle_{t-1})\right]^T V''(\tilde{\triangle}_{t-1})\gamma_t(z^0+\triangle_{t-1})\Psi_t(z^0+\triangle_{t-1}),\end{aligned}$$

where $\tilde{\triangle}_{t-1} \in \mathbb{R}^m$ is $\mathcal{F}_{t-1}$-measurable. Using (2.2) and (2.3) and taking the conditional expectation w.r.t. $\mathcal{F}_{t-1}$ yields

$$E\left\{V(\triangle_t) \mid \mathcal{F}_{t-1}\right\} \leq V(\triangle_{t-1}) + \mathcal{N}_t(\triangle_{t-1}). \tag{2.6}$$

Using the obvious decomposition $\mathcal{N}_t(\triangle_{t-1}) = \left[\mathcal{N}_t(\triangle_{t-1})\right]^+ - \left[\mathcal{N}_t(\triangle_{t-1})\right]^-$, we can write

$$\begin{aligned}\mathcal{N}_t(\triangle_{t-1}) &= (1+V(\triangle_{t-1}))^{-1}\left[\mathcal{N}_t(\triangle_{t-1})\right]^+(1+V(\triangle_{t-1})) - \left[\mathcal{N}_t(\triangle_{t-1})\right]^-\\ &= B_t\,(1+V(\triangle_{t-1})) - \left[\mathcal{N}_t(\triangle_{t-1})\right]^-.\end{aligned}$$

where

$$B_t = (1+V(\triangle_{t-1}))^{-1}\left[\mathcal{N}_t(\triangle_{t-1})\right]^+.$$

Hence (2.6) implies that

$$E\left\{V(\triangle_t) \mid \mathcal{F}_{t-1}\right\} \leq V(\triangle_{t-1})(1+B_t) + B_t - \left[\mathcal{N}_t(\triangle_{t-1})\right]^-, \tag{2.7}$$

eventually and, by (2.4),

$$\sum_{t=1}^{\infty} B_t < \infty. \tag{2.8}$$

According to the Robbins-Siegmund Lemma (see e.g., [22]) inequalities (2.7) and (2.8) imply that (2.5) holds and $V(\triangle_t)$ converges to some finite limit. $\diamondsuit$

Remark 2.3 *To describe the meaning of the conditions, let us consider a one dimensional case and assume for simplicity that the step-size sequence is state independent and positive, i.e.* $\gamma_t(u) = \gamma_t > 0$. *Assume also that* $V(u) = u^2$, *which is the most common choice of the function* V. *Then, the definition of the truncation operator ensures that condition* **(L)** *holds. Also, since* $V'(u) = 2u$ *and* $V''(u) = 2$,

$$\mathcal{N}_t(u) = 2u\gamma_t R_t(z^0 + u) + \gamma_t^2 E\left\{\Psi_t^2(z^0 + u) \mid \mathcal{F}_{t-1}\right\} \tag{2.9}$$

and since $E\left\{\varepsilon_t(z) \mid \mathcal{F}_{t-1}\right\} = 0$ *and* $R_t(z)$ *is* $\mathcal{F}_{t-1}$*-measurable, the second term of* $\mathcal{N}$ *can be written as*

$$\gamma_t^2 E\left\{\Psi_t^2(z^0 + u) \mid \mathcal{F}_{t-1}\right\} = \gamma_t^2 R_t^2(z^0 + u) + \gamma_t^2 E\left\{\varepsilon_t^2(z^0 + u) \mid \mathcal{F}_{t-1}\right\}. \tag{2.10}$$

Now recall that a typical assumption in SA is that the derivative of R *function at the root* z^0 *is negative. So, the first term of* $\mathcal{N}$ *is expected to be negative at least for small values of* u. *It therefore follows that in* (2.4), $[\mathcal{N}_t(u)]^+$ *can be replaced by its second term* (2.10) *implying that* (2.4) *holds if*

$$\sum_{t=1}^{\infty} \gamma_t^2 \, \frac{R_t^2(\triangle_{t-1})}{1 + \triangle_{t-1}^2} < \infty \quad \textit{and} \quad \sum_{t=1}^{\infty} \gamma_t^2 \, \frac{E\left\{\varepsilon_t(\triangle_{t-1})^2 \mid \mathcal{F}_{t-1}\right\}}{1 + \triangle_{t-1}^2} < \infty \tag{2.11}$$

Assuming that $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$ *, the two conditions above are very similar to the classical ones in SA. Recall that in the classical case when* $R_t(u) = R(u)$, *a standard assumption is that* $R^2(u) \le B(1 + u^2)$ *for some positive* B. *So, the first condition in* (2.11) *restricts the rate of growth of* R *functions w.r.t.* u *at infinity. The second part of* (2.11) *is also a natural generalisation of the corresponding condition in the classical SA. If for example, the error terms are state independent, then it reduces to* (2.17) *below. This implies that the variances (or conditional variances) of the error terms can even go to infinity, as far as the finiteness of the sum in* (2.17) *holds. It also follows from the above analysis that the step-size sequence can go to zero at any rate as far as* $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$. *Note that Theorem 2.2 does not assert convergence of the procedure to the root* z^0. *It establishes a stability type result stating that* $|Z_t - z^0|$ *converges to a finite limit. Note also that in Theorem 2.2, a rapidly decreasing step-size sequence is preferred. However, to have the convergence of the procedure to the root* z^0, *one must ensure that the convergence of* γ_t *is not too fast, similarly to the case of the classical SA. The requirements in Theorem 2.4 below put a certain limit to the rate at which the step-size sequence decreases to zero (see also Remark 2.5 below).*

Everywhere below, we assume that the $\inf_{u\in U} v(u)$ of a function $v(u)$ is 1 whenever $U = \emptyset$.

Theorem 2.4 *Suppose that $V(Z_t - z^0)$ converges (P-a.s.) to a finite limit for any initial value Z_0, where Z_t and V are defined in Theorem 2.2, and (2.5) holds. Suppose also that for each $\varepsilon \in (0,1)$,*

$$\inf_{\substack{\|u\|\geq\varepsilon \\ z^0+u\in U_t}} V(u) > \delta > 0 \tag{2.12}$$

eventually, for some δ. Suppose also that

(C) *For each $\varepsilon \in (0,1)$,*

$$\sum_{t=1}^{\infty} \inf_u [\mathcal{N}_t(u)]^- = \infty, \qquad P\text{-a.s.}$$

where the infimum is taken over the set $\{u : \varepsilon \leq V(u) \leq 1/\varepsilon;\ z^0 + u \in U_{t-1}\}$.

Then $Z_t \to z^0$ (P-a.s.), for any initial value Z_0.

Proof. As always (see the convention in 2.1), convergence and all relations between random variables are meant with probability one w.r.t. the measure P unless specified otherwise. Suppose that $V(\triangle_t) \to r \geq 0$ and there exists a set A with $P(A) > 0$, such that $r > 0$ on A. Then there exists $\varepsilon > 0$ and (possibly random) t_0, such that if $t \geq t_0$, $\varepsilon \leq V(\triangle_{t-1}) \leq 1/\varepsilon$ on A. Note also that $z^0 + \triangle_{t-1} = Z_{t-1} \in U_{t-1}$. By **(C)**, these would imply that

$$\sum_{s=t_0}^{\infty} [\mathcal{N}_s(\triangle_{s-1})]^- \geq \sum_{s=t_0}^{\infty} \inf_u [\mathcal{N}_s(u)]^- = \infty$$

on the set A, where the infimums are taken over the sets specified in condition (**C**). This contradicts (2.5). Hence, $r = 0$ and so, $V(\triangle_t) \to 0$. Now, $\triangle_t \to 0$ follows from (2.12) by contradiction. Indeed, suppose that $\triangle_t \not\to 0$ on a set, say B of positive probability. Then, for any fixed ω from this set, there would exist a sequence $t_k \to \infty$ such that $\|\triangle_{t_k}\| \geq \varepsilon$ for some $\varepsilon > 0$, and (2.12) would imply that $V(\triangle_{t_k}) > \delta > 0$ for large k-s, which contradicts the P-a.s. convergence $V(\triangle_t) \to 0$. $\diamondsuit$

Remark 2.5 *As in Remark 2.3, let us assume that $V(u) = u^2$ and the step-size sequence is state independent and positive. Then the $\mathcal{N}_t$ sequence can be written as (2.9). Condition* **(S)** *in Theorem 2.2 ensures that the second term in (2.9) is small (see Remark 2.3). To have the convergence of the procedure to the root z^0, condition* **(C)** *in Theorem 2.4 ensures that the first term, which should be negative, does not vanish too rapidly. For the classical SA, with a smooth R function having the property that $uR(z^0 + u) < 0$, this condition holds if $\sum_{t=1}^{\infty} \gamma_t = \infty$. In general, one must ensure that the derivatives of the R_t functions at z^0 do not decrease in absolute value too rapidly as t goes to infinity.*

2.3 Sufficient conditions

Remark 2.6 *Convergence results in the theorems of the previous section are global in the sense that they establish convergence of procedure* (2.1) *for any starting point. The role of the truncation operator, which is not immediately evident, can be seen by noting that the conditions in these theorems are formulated along the sequence* $\triangle_t = Z_t - z^0$. *Since for each* t, Z_t *belongs to the truncation set* U_t, *sufficient conditions can be written taking into account that the arguments of the corresponding functions belong to* U_t. *This can weaken some requirements considerably, if for example* U_t *is bounded (see e.g., examples 1 and 2 below).*

Everywhere in this subsection we assume that γ_t *is state independent (i.e., constant w.r.t.* z*) non-negative scalar predictable process.*

Corollary 2.7 *Let* Z_t *be a process defined by* (2.1), (2.2) *and* (2.3), *with an admissible for* $z^0 \in \mathbb{R}^m$ *truncation sequence* U_t. *Suppose also that* γ_t *is a non-negative predictable scalar process and*

(C1)

$$\sup_{z \in U_{t-1}} \frac{\left[2(z-z^0)^T R_t(z) + \gamma_t E\left\{\|\Psi_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}\right]^+}{1+\|z-z^0\|^2} \le q_t \tag{2.13}$$

eventually, where

$$\sum_{t=1}^{\infty} q_t \gamma_t < \infty, \qquad P\text{-a.s.}$$

Then $\|Z_t - z^0\|$ *converges (*P*-a.s.) to a finite limit.*

Proof. Let us show that the conditions of Theorem 2.2 are satisfied with $V(u) = u^T u = \|u\|^2$ and the step-size sequence $\gamma_t(z) = \gamma_t \mathbf{I}$. Since $z^0 \in U_t$ for large t-s, the definition of the truncation (see 2.1) implies that

$$\|Z_t - z^0\| \le \left\|Z_{t-1} + \gamma_t \Psi_t(Z_{t-1}) - z^0\right\|,$$

eventually. Therefore (**L**) holds. Then, $V'(u) = 2u^T$ and $V''(u) = 2\mathbf{I}$, and so, for the process $\mathcal{N}_t(u)$ in (2.4) we have

$$\mathcal{N}_t(u) = 2u^T \gamma_t R_t(z^0+u) + \gamma_t^2 E\left\{\|\Psi_t(z^0+u)\|^2 \mid \mathcal{F}_{t-1}\right\} \tag{2.14}$$

and

$$\frac{[\mathcal{N}_t(\triangle_{t-1})]^+}{1+V(\triangle_{t-1})} = \gamma_t \, \frac{\left[2\triangle_{t-1}^T R_t(z^0+\triangle_{t-1}) + \gamma_t E\left\{\|\Psi_t(z^0+\triangle_{t-1})\|^2 \mid \mathcal{F}_{t-1}\right\}\right]^+}{1+\|\triangle_{t-1}\|^2}$$

Since $z^0 + \triangle_{t-1} = Z_{t-1} \in U_{t-1}$, (2.4) follows from (**C1**). $\diamondsuit$

Corollary 2.8 *Suppose that the conditions of Corollary 2.7 hold and*

(C2) *for each* $\varepsilon \in (0,1)$,

$$\sum_{t=1}^{\infty} \inf_{u} [\mathcal{N}_t(u)]^- = \infty, \qquad P\text{-a.s.}$$

where

$$\mathcal{N}_t(u) = 2u^T \gamma_t R_t(z^0+u) + \gamma_t^2 E\left\{\|\Psi_t(z^0+u)\|^2 \mid \mathcal{F}_{t-1}\right\}$$

and the infimum is taken over the set $\{u : \varepsilon \le \|u\| \le 1/\varepsilon;\ z^0+u \in U_{t-1}\}$. *Then* $Z_t \to z^0$ *(P-a.s.), for any initial value* Z_0.

Proof. Let us show that the conditions of Theorem 2.4 are satisfied with $V(u) = u^T u = \|u\|^2$ and $\gamma_t(z) = \gamma_t \mathbf{I}$. It follows from the proof of Corollary 2.7 that all the conditions of Theorem 2.2 hold with $V(u) = u^T u$. Hence, $\|Z_t - z^0\|$ converges and (2.5) holds. Since

$$\inf_{\substack{\|u\| \ge \varepsilon \\ z^0+u \in U_t}} \|u\|^2 \ge \varepsilon^2,$$

condition (2.12) also trivially holds. Finally, (**C**) is a consequence of (**C2**). $\diamondsuit$

Remark 2.9 *The corollaries below show that, under the condition* $(z-z^0)R_t(z) < 0$, *convergence of* $|Z_t - z^0|$ *to a finite limit is determined by the statistical properties of the error terms, and the behaviour of the* R_t *functions at the points that are far away from* z^0 *(see Corollary 2.10 below). On the other hand, convergence to the root* z^0 *is largely determined by the local properties of the* R_t *functions at* z^0.

Corollary 2.10 *Suppose that* Z_t *is a process defined by* (2.1), (2.2) *and* (2.3), *with an admissible for* $z^0 \in \mathbb{R}^m$ *truncation sequence* U_t *and*

(1)

$$(z-z^0)^T R_t(z) \le 0 \qquad \textit{for any} \qquad z \in U_t,$$

eventually;

(2)

$$\sup_{z \in U_{t-1}} \frac{\|R_t(z)\|^2}{1+\|z-z^0\|^2} \le r_t$$

eventually, where

$$\sum_{t=1}^{\infty} r_t \gamma_t^2 < \infty, \qquad P\text{-a.s.},$$

(3)

$$\sup_{z\in U_{t-1}} \frac{E\left\{\|\varepsilon_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}}{1+\|z-z^0\|^2} \le e_t$$

eventually, where

$$\sum_{t=1}^{\infty} e_t\gamma_t^2 < \infty, \qquad P\text{-a.s.}.$$

Then $\|Z_t - z^0\|$ converges (P-a.s.) to a finite limit.

Proof. Using condition (1),

$$\left[2(z-z^0)^T R_t(z) + \gamma_t E\left\{\|\Psi_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}\right]^+ \le \gamma_t E\left\{\|\Psi_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}$$

eventually. Since $E\left\{\varepsilon_t(z) \mid \mathcal{F}_{t-1}\right\} = 0$ and $R_t(z)$ is $\mathcal{F}_{t-1}$-measurable, we have

$$E\left\{\|\Psi_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\} = \|R_t(z)\|^2 + E\left\{\|\varepsilon_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}. \tag{2.15}$$

So, by conditions (2) and (3), the left hand side of (2.13) does not exceed $(r_t+e_t)\gamma_t$. Hence conditions of Corollary 2.7 hold with $q_t = (r_t+e_t)\gamma_t$ and the result follows. ♢

Corollary 2.11 *Suppose that the conditions of Corollary 2.10 are satisfied and*

(CC) *for each $\varepsilon \in (0,1)$,*

$$\inf_{\substack{\varepsilon\le\|z-z^0\|\le 1/\varepsilon \\ z\in U_{t-1}}} -(z-z^0)^T R_t(z) > \nu_t \tag{2.16}$$

eventually, where

$$\sum_{t=1}^{\infty} \nu_t\gamma_t = \infty, \qquad P\text{-a.s.}$$

Then Z_t converges (P-a.s.) to z^0.

Proof. It follows from the poof of Corollary 2.10 that conditions of Corollary 2.7 hold. Let us prove that (**C2**) of Corollary 2.8 holds. Using the obvious inequality $[a]^- \ge -a$, we have

$$[\mathcal{N}_t(u)]^- \ge -2u^T\gamma_t R(z^0+u) - \gamma_t^2 E\left\{\|\Psi_t(z^0+u)\|^2 \mid \mathcal{F}_{t-1}\right\}.$$

Using (2.15) and conditions (2) and (3) of Corollary 2.10, and taking the supremum of the conditional expectation above over the set $\{u : \varepsilon \le \|u\| \le 1/\varepsilon;\ z^0+u \in U_{t-1}\}$, we obtain

$$\sup \frac{E\left\{\|\Psi_t(z^0+u)\|^2 \mid \mathcal{F}_{t-1}\right\}}{1+\|u\|^2}\ (1+\|u\|^2) \le (r_t+e_t)(1+\|1/\varepsilon\|^2).$$

Then, by (2.16), taking the infimum over the same set,

$$\inf [\mathcal{N}_t(u)]^- \geq 2\gamma_t\nu_t - \gamma_t^2(r_t + e_t)(1 + \|1/\varepsilon\|^2).$$

Condition (**C2**) is now immediate from (**CC**) and conditions (2) and (3) of Corollary 2.10. Hence, by Corollary 2.8, Z_t converges (P-a.s.) to z^0. $\diamondsuit$

Remark 2.12 *Suppose that ε_t is an error term which does not depend on z and denote*

$$\sigma_t^2 = E\left\{\|\varepsilon_t\|^2 \mid \mathcal{F}_{t-1}\right\}$$

Then condition (3) holds if

$$\sum_{t=1}^{\infty} \sigma_t^2\gamma_t^2 < \infty, \qquad P\text{-a.s.}. \tag{2.17}$$

This shows that the requirement on the error terms are quite weak. In particular, the conditional variances do not have to be bounded w.r.t. t.

Remark 2.13 *As it was mentioned in the introduction, our procedure is similar to the one considered in [1]. Let us compare these two in the cases when the comparisons are possible. Hence, consider truncations on increasing non-random sets, non-random and homogeneous $R_t(u) = R(u)$, and scalar and state-independent γ_t in Corollaries 2.10 and 2.11. Also, in Theorem 2 of [1] take $\beta_n = 0$. Then the resulting two sets of conditions are in fact equivalent. In particular, in terms of notation in [1],*

$$a_n = \gamma_n, \quad \frac{1}{c_n^2} = e_n, \quad M_n^2 = r_n.$$

*Now it is clear that conditions 2. and 3. in Theorem 2 of [1] are equivalent to (3) and (2) respectively in Corollary 2.10. Note that although condition (**CC**) in 2.11 is formally more general than Condition 2 in Theorem 2 of [1], in any meaningful applications they are equivalent.*

2.4 Examples

Example 1 Let l be an odd integer and

$$R(z) = -(z - z^0)^l,$$

$z, z^0 \in \mathbb{R}$. Consider a truncation sequence $[-\alpha_t, \alpha_t]$, where $\alpha_t \to \infty$ is a sequence of positive numbers. Suppose that

$$\sum_{t=1}^{\infty} \gamma_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \alpha_{t-1}^{2l}\ \gamma_t^2 < \infty.$$

Then, provided that the measurement errors satisfy (2.17) (or condition (3) of Corollary 2.10 in the case of state-dependent errors), the truncated procedure

$$Z_t = \left[\ Z_{t-1} + \gamma_t \left(R(Z_{t-1}) + \varepsilon_t\right)\right]_{-\alpha_t}^{\alpha_t}, \quad t = 1, 2, \ldots$$

converges a.s. to z^0.

Indeed, condition (1) of Corollary 2.10 trivially holds. For large t's,

$$\sup_{z \in [-\alpha_{t-1}, \alpha_{t-1}]} \frac{\|R(z)\|^2}{1 + \|z - z^0\|^2} \le \sup_{z \in [-\alpha_{t-1}, \alpha_{t-1}]} (z - z^0)^{2l} \le 4^l \alpha_{t-1}^{2l}$$

which implies condition (2) of Corollary 2.10. Condition (**CC**) of Corollary 2.11 also trivially holds with $\nu_t = \varepsilon^{l+1}$.

For example, if the degree of the polynomial is known to be l (or at most l), and $\gamma_t = 1/t$, then one can take $\alpha_t = Ct^{\frac{1}{2l} - \delta}$, where C and δ are some positive constants and $\delta < \frac{1}{2l}$. One can also take a truncation sequence which is independent of l, e.g., $\alpha_t = C \log t$, where C is a positive constant.

Example 2 Let $X_1, X_2, \ldots,$ be i.i.d. Gamma$(\theta, 1)$, $\theta > 0$. Then the the common probability density function is

$$f(x, \theta) = \frac{1}{\mathbf{\Gamma}(\theta)} x^{\theta - 1} e^{-x}, \quad \theta > 0, \ \ x > 0,$$

where $\mathbf{\Gamma}(\theta)$ is the Gamma function. Then

$$\frac{f'(x, \theta)}{f(x, \theta)} = \log x - \underbrace{\frac{d}{d\theta} \log \mathbf{\Gamma}(\theta)}_{\log' \mathbf{\Gamma}(\theta)}, \qquad i(\theta) = \underbrace{\frac{d^2}{d\theta^2} \log \mathbf{\Gamma}(\theta)}_{\log'' \mathbf{\Gamma}(\theta)},$$

where $i(\theta)$ is the one-step Fisher information. Then a likelihood type recursive estimation procedure (see also (1.2)) can be defined as

$$\hat{\theta}_t = \left[\hat{\theta}_{t-1} + \frac{1}{t \, \log'' \mathbf{\Gamma}(\hat{\theta}_{t-1})} \left(\log X_t - \log' \mathbf{\Gamma}(\hat{\theta}_{t-1})\right)\right]_{\alpha_t}^{\beta_t}, \qquad t = 1, 2, \ldots \tag{2.18}$$

where $\alpha_t \downarrow 0$ and $\beta_t \uparrow \infty$ are sequences of positive numbers.

Everywhere in this example, $\mathcal{F}_t$ is the sigma algebra generated by $X_1, \ldots, X_t$, P^θ is the family of corresponding measures, and $\theta > 0$ is an arbitrary but fixed value of the parameter.

Let us rewrite (2.18) in the form of the stochastic approximation, i.e.,

$$\hat{\theta}_t = \left[\hat{\theta}_{t-1} + \frac{1}{t} \left(R(\hat{\theta}_{t-1}) + \varepsilon_t(\hat{\theta}_{t-1})\right)\right]_{\alpha_t}^{\beta_t}, \qquad t = 1, 2, \ldots \tag{2.19}$$

where (see Section 4 for details)

$$R(u) = R^{\theta}(u) = \frac{1}{\log''\mathbf{\Gamma}(u)} E^{\theta}\{\ln X_t - \log'\mathbf{\Gamma}(u)\} = \frac{1}{\log''\mathbf{\Gamma}(u)} \left(\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u)\right)$$

and

$$\varepsilon_t(u) = \frac{1}{\log''\mathbf{\Gamma}(u)} \left(\log X_t - \log'\mathbf{\Gamma}(u)\right) - R(u).$$

Since $E^{\theta}\left\{\log X_t \mid \mathcal{F}_{t-1}\right\} = E^{\theta}\left\{\log X_t\right\} = \log'\mathbf{\Gamma}(\theta)$ and $\hat{\theta}_{t-1}$ is $\mathcal{F}_{t-1}$ - measurable, we have $E^{\theta}\left\{\varepsilon_t(\hat{\theta}_{t-1}) \mid \mathcal{F}_{t-1}\right\} = 0$ and hence (2.2) holds. Since $E^{\theta}\left\{\log^2 X_t\right\} < \infty$, condition (2.3) can be checked in the similar way. Obviously, $R(\theta) = 0$, and since $\log'\mathbf{\Gamma}$ is increasing (see, e.g., [33], 12.16), condition **(1)** of Corollary 2.10 holds with $z^0 = \theta$. Based on the well known properties of the logarithmic derivatives of the gamma function, it is not difficult to show (see Section 4) that if

$$\sum_{t=1}^{\infty} \frac{\alpha_{t-1}^2}{t} = \infty \qquad \text{and} \qquad \sum_{t=1}^{\infty} \frac{\log^2 \alpha_{t-1} + \log^2 \beta_{t-1}}{t^2} < \infty, \tag{2.20}$$

then all the conditions of Corollary 2.10 and 2.11 hold and therefore, $\hat{\theta}_t$ is consistent, i.e.,

$$\hat{\theta}_t \to \theta \qquad \text{as} \qquad t \to \infty \quad (P^{\theta}\text{-a.s.}).$$

For instance, the sequences

$$\alpha_t = C_1(\log\,(t+2))^{-\frac{1}{2}} \quad \text{and} \quad \beta_t = C_2(t+2)$$

with some positive constants C_1 and C_2, obviously satisfy (2.20).

Note also, that since $\theta \in (0, \infty)$, it may seem unnecessary to use the upper truncations $\beta_t < \infty$. However, without upper truncations (i.e. if $\beta_t = \infty$), the standard restriction on the growth does not hold. Also, with $\beta_t = \infty$ the procedure fails condition (2) of Corollary 2.10 (see (4.7)).

Example 3 Consider an AR(1) process

$$X_t = \theta X_{t-1} + \xi_t, \tag{2.21}$$

where ξ_t is a sequence of random variables with mean zero. Taking

$$\Psi_t(z) = X_{t-1}\left(X_t - zX_{t-1}\right)$$

$\gamma_t(z) = \gamma_t = \hat{I}_t = \hat{I}_0 + \sum_{s=1}^{t} X_{t-1}^2$, and $U_t = \mathbb{R}$, procedure (2.1) reduces to the recursive least squares (LS) estimator of θ, i.e.,

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + \hat{I}_t^{-1} X_{t-1}\left(X_t - \hat{\theta}_{t-1}X_{t-1}\right), \\ \hat{I}_t &= \hat{I}_{t-1} + X_{t-1}^2, \qquad t = 1, 2, \ldots \end{aligned} \tag{2.22}$$

where $\hat{\theta}_0$ and $\hat{I}_0 > 0$ are any starting points.

For simplicity let us assume that ξ_t is a sequence of i.i.d. r.v.'s with mean zero and variance 1. Consistency of (2.22) can be derived from our results for any $\theta \in \mathbb{R}$ and without any further moment assumptions on the innovation process ξ_t. Indeed, assume that θ is an arbitrary but fixed value of the parameter. Then, using (2.21), we obtain

$$X_t - \hat{\theta}_{t-1} X_{t-1} = \xi_t + X_{t-1}(\theta - \hat{\theta}_{t-1}).$$

and (2.22) can be rewritten as

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \hat{I}_t^{-1} \left(X_{t-1}^2 (\theta - \hat{\theta}_{t-1}) + X_{t-1}\xi_t \right). \tag{2.23}$$

So, (2.23) is a SA procedure with

$$R_t(z) = X_{t-1}^2(\theta - z), \tag{2.24}$$

$\varepsilon_t(z) = \varepsilon_t = X_{t-1}\xi_t, \quad \gamma_t = \hat{I}_t^{-1}$ and $U_t = \mathbb{R}$. Let us check condition (**C1**) of Corrolary 2.7 with $z^0 = \theta$ and $U_t = \mathbb{R}$. Since $E\{\varepsilon_t \mid \mathcal{F}_{t-1}\} = 0$ and $R_t(z)$ is $\mathcal{F}_{t-1}$ measurable, (2.2) and (2.3) trivially hold. Also,

$$E\left\{\|\Psi_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\} = \|R_t(z)\|^2 + E\left\{\|\varepsilon_t\|^2 \mid \mathcal{F}_{t-1}\right\} = X_{t-1}^4(\theta - z)^2 + X_{t-1}^2, \tag{2.25}$$

denoting the expression in the square brackets in (2.13) by $w_t(z)$ (with $z^0 = \theta$),we obtain

$$w_t(z) = -2X_{t-1}^2(z-\theta)^2 + \hat{I}_t^{-1} X_{t-1}^4 (\theta - z)^2 + \hat{I}_t^{-1} X_{t-1}^2 \tag{2.26}$$

$$= -\delta X_{t-1}^2 (z-\theta)^2 - X_{t-1}^2 (z-\theta)^2 \left((2-\delta) - \hat{I}_t^{-1} X_{t-1}^2 \right) + \hat{I}_t^{-1} X_{t-1}^2 \tag{2.27}$$

for some $0 < \delta < 1$. Since $\hat{I}_t^{-1} X_{t-1}^2 \le 1$, the positive part of the above expression does not exceed $\hat{I}_t^{-1} X_{t-1}^2$. This implies that (2.13) holds with $q_t = \hat{I}_t^{-1} X_{t-1}^2$. Now, note that if d_n is a nondecreasing sequence of positive numbers such that $d_t \to +\infty$ and $\triangle d_t = d_t - d_{t-1}$, then $\sum_{t=1}^{\infty} \triangle d_t / d_t = +\infty$ and $\sum_{t=1}^{\infty} \triangle d_t / d_t^2 < +\infty$. So, for $X_{t-1}^2 = \triangle \hat{I}_t$, since $\hat{I}_t \to \infty$ for any $\theta \in \mathbb{R}$ (see, e.g, Shiryayev [30], Ch.VII, §5) , we have

$$\sum_{t=1}^{\infty} \hat{I}_t^{-2} X_{t-1}^2 < \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \hat{I}_t^{-1} X_{t-1}^2 = \infty. \tag{2.28}$$

Hence, taking $q_t \gamma_t = \hat{I}_t^{-2} X_{t-1}^2$, (**C1**) follows. Therefore, $(\hat{\theta}_t - \theta)^2$ converges to a finite limit. To show convergence to θ, let us check condition (**C2**) of of Corrolary 2.8 with $z^0 = \theta$ and $U_t = \mathbb{R}$. Using (2.24) and (2.25), we have

$$\mathcal{N}_t(u) = -2\hat{I}_t^{-1} X_{t-1}^2 u^2 + \hat{I}_t^{-2} X_{t-1}^4 u^2 + \hat{I}_t^{-2} X_{t-1}^2 = \hat{I}_t^{-1} w_t(\theta + u),$$

where w_t is defined in (2.26). Since the middle term in (2.27) is non-positive, using the obvious inequality $[a]^- \ge -a$, we can write

$$[\mathcal{N}_t(u)]^- \ge \delta \hat{I}_t^{-1} X_{t-1}^2 u^2 - \hat{I}_t^{-2} X_{t-1}^2,$$

and

$$\sum_{t=1}^{\infty} \inf_{\varepsilon \le |u| \le 1/\varepsilon} [\mathcal{N}_t(u)]^- = \infty$$

now follows from (2.28). So, by Corollary 2.7, $\hat{\theta}_t \to \theta$ ($P^\theta-$ a.s.).

Note that the convergence of the LS estimator is well known under these assumptions. (see e.g., [30], Ch.VII, §5). This example is presented to demonstrate that the assumptions made here are minimal. That is, in well know model cases, the results of the paper do not assume any additional restrictions.

3 Concluding remarks

The paper establishes convergence of SA procedures with the following features:

- inhomogeneous random functions R_t,
- state dependent matrix valued random step sizes,
- truncations with random and moving (shrinking or expanding) bounds.

Conditions introduced in the paper can be divided into two main groups. The first group is concerned with statistical properties of the error terms, and the behaviour of the R_t functions at the points that are far away from the root z^0. They guarantee a stability type property of the SA procedure ensuring that $\|Z_t - z^0\|$ converges to a finite limit. The second group of conditions is mostly concerned with local properties of the R_t functions at z^0. These conditions ensure convergence of the procedure to the root z^0. While in the first group, a rapidly decreasing step-size sequence is preferred, the second group puts a certain limit to the rate at which the step-size sequence decreases to zero.

4 Appendix

We will need the following properties of the Gamma function (see, e.g., [33], 12.16). $\log'\mathbf{\Gamma}$ is increasing, $\log''\mathbf{\Gamma}$ is decreasing and continuous, and

$$\log''\mathbf{\Gamma}(x) = \frac{1}{x^2} + \sum_{n=1}^{\infty} \frac{1}{(x+n)^2}.$$

The latter implies that

$$\log''\mathbf{\Gamma}(x) \le \frac{1}{x^2} + \sum_{n=1}^{\infty} \int_{n-1}^{n} \frac{dz}{(x+z)^2} = \frac{1}{x^2} + \frac{1}{x} = \frac{1+x}{x^2} \tag{4.1}$$

and

$$\log''\mathbf{\Gamma}(x) \geq \sum_{n=0}^{\infty} \int_n^{n+1} \frac{dz}{(x+z)^2} = \frac{1}{x}. \tag{4.2}$$

Also (see [8], 12.5.4),

$$\log'\mathbf{\Gamma}(x) \leq \ln(x). \tag{4.3}$$

Then,

$$E^\theta \{\log X_1\} = \log'\mathbf{\Gamma}(\theta) \quad \text{and} \quad E^\theta \left\{(\log X_1)^2\right\} = \log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta))^2 \tag{4.4}$$

and

$$E^\theta \left\{(\log X_1 - \log'\mathbf{\Gamma}(\theta))^2\right\} = \log''\mathbf{\Gamma}(\theta).$$

Let us show that the conditions of Corollary 2.10 hold. Since

$$\Psi_t(u) = \frac{1}{\log''\mathbf{\Gamma}(u)} \left(\log X_t - \log'\mathbf{\Gamma}(u)\right),$$

using (4.4) and (4.2) we obtain

$$\begin{aligned} \frac{E\left\{\|\Psi_t(u)\|^2 \mid \mathcal{F}_{t-1}\right\}}{1+\|u-\theta\|^2} &= \frac{\log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u))^2}{(\log''\mathbf{\Gamma}(u))^2(1+\|u-\theta\|^2)} \\ &\leq \frac{u^2}{1+(u-\theta)^2} \left(\log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u))^2\right). \end{aligned} \tag{4.5}$$

Now, $u^2/(1+(u-\theta)^2) \leq C$. Here and further on in this subsection, C denotes various constants which may depend on θ. So, using (4.3) we obtain

$$\frac{E\left\{\|\Psi_t(u)\|^2 \mid \mathcal{F}_{t-1}\right\}}{1+\|u-\theta\|^2} \leq C\left(\log''\mathbf{\Gamma}(\theta) + \log'\mathbf{\Gamma}(\theta)^2 + \log'\mathbf{\Gamma}(u)^2\right) \leq C(1+\log^2(u)).$$

For large t's, since $\alpha_t < 1 < \beta_t$, we have

$$\sup_{u\in[\alpha_t,\beta_t]} \log^2(u) \leq \left\{ \sup_{\alpha_t \leq u < 1} \log^2(u) + \sup_{1<u\leq\beta_t} \log^2(u) \right\} \leq \log^2\alpha_t + \log^2\beta_t.$$

Condition (2) of Corollary 2.10 is now immediate from the second part of (2.20). It remains to check that (**CC**) of Corollary 2.11 holds. Indeed,

$$-(u-\theta)R(u) = \frac{(u-\theta)\left(\log'\mathbf{\Gamma}(u) - \log'\mathbf{\Gamma}(\theta)\right)}{\log''\mathbf{\Gamma}(u)}.$$

Since $\log'\mathbf{\Gamma}$ is increasing and $\log''\mathbf{\Gamma}$ is decreasing and continuous, we have that for each $\varepsilon \in (0,1)$,

$$\inf_{\substack{\varepsilon\leq\|u-\theta\|\leq 1/\varepsilon \\ u\in U_{t-1}}} -(u-\theta)R(u) \geq \frac{\inf_{\varepsilon\leq\|u-\theta\|\leq 1/\varepsilon}\left(\log'\mathbf{\Gamma}(u) - \log'\mathbf{\Gamma}(\theta)\right)(u-\theta)}{\sup_{u\in U_{t-1}} \log''\mathbf{\Gamma}(u)} \geq \frac{C}{\log''\mathbf{\Gamma}(\alpha_{t-1})} \tag{4.6}$$

where C is a constant that my depend on ε and θ. Since $\alpha_{t-1} < 1$ for large t's, it follows (4.1) that $1/\log''\mathbf{\Gamma}(\alpha_{t-1}) \geq \alpha_{t-1}^2/2$. Condition (**CC**) of Corollary 2.11 is now immediate from the first part of (2.20).

Note that with $\beta_t = \infty$ the procedure fails condition (2) of Corollary 2.10. Indeed, (4.5) and (4.1) implies that

$$\sup_{\alpha_t \leq u} \frac{E\left\{\Psi_t^2(u) \mid \mathcal{F}_{t-1}\right\}}{1+(u-\theta)^2} \geq \sup_{\alpha_t \leq u} \frac{\left\{\log''\mathbf{\Gamma}(\theta) + \left(\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u)\right)^2\right\} u^4}{(1+u)^2(1+(u-\theta)^2)} = \infty \quad (4.7)$$

References

[1] Andradóttir, S. (1995). A stochastic approximation algorithm with varying bounds. *Operations Research* **43**, 6, 1037–1048.

[2] Andrieu, C., Moulines, E. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44**, 283–312.

[3] Benveniste, A, Metivier, M. and Priouret , P. (1990). *Adaptive Algorithms and Stochastic Approximation*. Berlin and New York: Springer-Verlag.

[4] Borkar, V. S. (2008). *Stochastic approximation: A Dynamical Systems Viewpoint.* Cambridge University Press.

[5] Chen, H., Guo, L. and Gao, A. (1987). Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Processes Appl.* **27**, 217231.

[6] Chen, H. and Zhu, Y.-M. (1986). Stochastic approximation procedures with randomly varying truncations. *Scientia Sinica 1* **29**, 914926.

[7] Campbell, K. (1982). Recursive computation of M-estimates for the parameters of a finite autoregressive process. *Ann. Statist.* **10**, 442-453.

[8] Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[9] Englund, J.-E., Holst, U., and Ruppert, D. (1989). Recursive estimators for stationary, strong mixing processes – a representation theorem and asymptotic distributions. *Stochastic Processes Appl.* **31**, 203–222.

[10] Fabian, V. (1978). On asymptotically efficient recursive estimation. *Ann. Statist.* **6**, 854-867.

[11] Gu, M.G. and Li, S. (1998). A stochastic approximation algorithm for maximum-likelihoood estimation with incomplete data. *The Canadian Journal of Statistics* **26**, 567-582.

[12] Kallenberg, O. (1997). *Foundations of Modern Probability.* Nauka, Moscow.

[13] Khas'minskii, R.Z., Nevelson, M.B. (1972). *Stochastic Approximation and Recursive Estimation.* Nauka, Moscow.

[14] Kushner, H. (2010). Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 6, 87–96.

[15] Kushner, H. and Yin, G. (1997). *Stochastic Approximation Algorithms and Applications. Applications of Mathematics.* Springer-Verlag, New-York.

[16] Lai, T.L. (2003). Stochastic approximation. *Ann. Statist.* **31**, 391-406.

[17] Lazrieva, N., Toronjadze, T, and Sharia, T. (2008). Semimartingale Stochastic Approximation Procedures. *Journal of Mathematical Sciences*, **153**, 3, 211 – 261.

[18] Lelong, J. (2008). Almost sure convergence of randomly truncated stochastic algorithms under verifiable conditions. *Statistics & probability Letters.* **28**, 2632-2636.

[19] Ljung, L. and Soderstrom, T. (1987). *Theory and Practice of Recursive Identification,* MIT Press.

[20] Polyak, B. T. and Tsypkin, Ya. Z. (1980). Robust identification. *Automatica* 16, 53–69

[21] Robbins, H. and Monro, S. (1951) A stochastic approximation method, *Ann. Statist.* **22**, 400–407.

[22] Robbins, H. and Siegmund, D. (1971). A convergence theorem for non-negative almost supermartingales and some applications. *Optimizing Methods in Statistics.* ed. J.S. Rustagi Academic Press, New York, 233–257.

[23] Sakrison, D.J. (1965). Efficient recursive estimation; application to estimating the parameters of a covariance function. *Internat. J. Engrg. Sci.* **3**, 461–483.

[24] Sharia, T. (1997). Truncated recursive estimation procedures, *Proc. A. Razmadze Math. Inst.* **115**, 149–159.

[25] Sharia, T. (1998). On the recursive parameter estimation for the general discrete time statistical model. *Stochastic Processes Appl.* **73**, **2**, 151–172.

[26] Sharia, T. (2008). Recursive parameter estimation: Convergence. *Statistical Inference for Stochastic Processes.* **11**, 2, pp. 157 – 175.

[27] Sharia, T. (2007). Rate of convergence in recursive parameter estimation procedures. *Georgian Mathematical Journal.* **14**, 4, pp. 721–736.

[28] Sharia, T. (2010). Recursive parameter estimation: Asymptotic expansion. *The Annals of The Institute of Statistical Mathematics* **62** 2, 343-362.

[29] Sharia, T. (2010). Efficient On-Line Estimation of Autoregressive Parameters. *Mathematical Methods of Statistics*. **19**, 2, 163-186.

[30] Shiryayev, A.N. (1984). *Probability,* Springer-Verlag, New York.

[31] Tadic, V. (1997) Stochastic gradient with random truncations, *European J. of Operational Research*, **101**, pp. 261–284.

[32] Tadic, V. (1998) Stochastic approximations with random truncations, state dependent noise and discontinuous dynamics, *Stochastics and Stochastics reports*. **64**, pp. 283–326.

[33] Whittaker, E. Watson, G. (1927). *A Course of Modern Analysis*. Cambridge University Press, Cambridge.