Rate of Convergence of Truncated Stochastic Approximation Procedures with Moving Bounds

Teo Sharia and Lei Zhong

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

Abstract

The paper is concerned with stochastic approximation procedures having three main characteristics: truncations with random moving bounds, a matrix valued random step-size sequence, and a dynamically changing random regression function. We study convergence and rate of convergence. Main results are supplemented with corollaries to establish various sets of sufficient conditions, with the main emphases on the parametric statistical estimation. The theory is illustrated by examples and special cases.

Keywords: Stochastic approximation, Recursive estimation, Parameter estimation

1 Introduction

This paper is a continuation of Sharia (2014) where a large class of truncated Stochastic approximation (SA) procedures with moving random bounds was proposed. Although the proposed class of procedures can be applied to a wider range of problems, our main motivation comes from applications to parametric statistical estimation theory. To make this paper self contained, we introduce the main ideas below (a full list of references as well as some comparisons can be found in Sharia (2014)).

The main idea can be easily explained in the case of 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

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

where $\{\varepsilon_t\}$ is a sequence of zero-mean random variables and $\{\gamma_t\}$ is a deterministic sequence of positive numbers. This is the classical Robbins-Monro SA procedure (see Robbins and Monro (1951)), which under certain conditions converges to the root z^0 of the equation $R(z) = 0$. (Comprehensive surveys of the SA technique can be found in Benveniste et al. (1990), Borkar (2008), Kushner and Yin (2003), Lai (2003), and Kushner (2010).)

Statistical parameter estimation is one of the most important applications of the above procedure. Indeed, suppose that $X_1, \ldots, X_t$ are i.i.d. random variables and $f(x, \theta)$ is the common probability density function (w.r.t. some σ-finite measure), where $\theta \in \mathbb{R}^m$ is an unknown parameter. Consider a recursive estimation procedure for θ defined by

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{t} i(\hat{\theta}_{t-1})^{-1} \, \frac{f'^T(X_t, \hat{\theta}_{t-1})}{f(X_t, \hat{\theta}_{t-1})}, \qquad t \geq 1, \tag{1.2}$$

where $\hat{\theta}_0 \in \mathbb{R}^m$ is some starting value and $i(\theta)$ is the one-step Fisher information matrix (f' is the row-vector of partial derivatives of f w.r.t. the components of θ). This estimator was introduced in Sakrison (1965) and studied by a number of authors (see e.g, Polyak and Tsypkin (1980), Campbell (1982), Ljung and Soderstrom (1987), Lazrieve and Toronjadze (1987), Englund et al (1989), Lazrieve et al (1997, 2008), Sharia (1997–2010)). 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. One can analyse (1.2) by rewriting it in the form of stochastic approximation with $\gamma_t = 1/t$,

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

where θ is an arbitrary but fixed value of the unknown parameter. Indeed, under certain standard assumptions, $R(\theta) = 0$ and $\{\varepsilon_t\}$ is a martingale difference w.r.t. the filtration $\{\mathcal{F}_t\}$ generated by $\{X_t\}$. So, (1.2) is a standard SA of type (1.1).

Suppose now that we have a stochastic process $X_1, X_2, \ldots$ and let $f_t(x, \theta) = f_t(x, \theta | X_1, \ldots, X_{t-1})$ be the conditional probability density function of the observation X_t given $X_1, \ldots, X_{t-1}$, where $\theta \in \mathbb{R}^m$ is an unknown parameter. Then one can define a recursive estimator of θ 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 $\psi_t(\theta) = \psi_t(X_1, \ldots, X_t; \theta)$, $t = 1, 2, \ldots$, are suitably chosen functions which may, in general, depend on the vector of all past and present observations $X_1, ..., X_t$, and have the property that the process $\psi_t(\theta)$ is P^θ- martingale difference, i.e., $E_\theta\left\{\psi_t(\theta) \mid \mathcal{F}_{t-1}\right\} = 0$ for each t. For example, a choice

$$\psi_t(\theta) = l_t(\theta) \equiv \frac{[f_t'(X_t,\theta)]^T}{f_t(X_t,\theta)}$$

yields a likelihood type estimation procedure. In general, to obtain an estimator with asymptotically optimal properties, a state-dependent matrix-valued random step-size sequences are needed (see Sharia (2010)). For the above procedure, a step-size sequence $\gamma_t(\theta)$ with the property

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

is an optimal choice. For example, to derive a recursive procedure which is asymptotically equivalent to the maximum likelihood estimator, we need to take

$$\psi_t(\theta) = l_t(\theta) \quad \text{and} \quad \gamma_t(\theta) = I_t^{-1}(\theta),$$

where

$$I_t(\theta) = \sum_{s=1}^{t} E\{l_s(\theta) l_s^T(\theta) | \mathcal{F}_{s-1}\} \tag{1.4}$$

is the conditional Fisher information matrix. To rewrite (1.3) in the SA form, let us assume that θ is an arbitrary but fixed value of the parameter and define

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

Then, since $\psi_t(\theta)$ is P^θ-martingale difference, it follows that $R_t(\theta) = 0$ for each t. So, the objective now is to find a common root θ of a dynamically changing sequence of functions R_t.

Before introducing the general SA process, let us consider one simple modification of the classical SA procedure. Suppose that we have additional information about the root z^0 of the equation $R(z) = 0$. Let us, e.g., assume that $z^0 \in [\alpha_t, \beta_t]$ at each step t, where α_t and β_t are random variables such that $-\infty < \alpha_t \le \beta_t < \infty$. Then one can consider a procedure, which at each step t produces points from the interval $[\alpha_t, \beta_t]$. For example, a truncated classical SA procedure in this case can be derived using the following recursion

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

where Φ is the truncation operator, that is, for any $-\infty < a \leq b < \infty$,

$$\Phi_{[a,b]}(z) = \begin{cases} a & \text{if } z < a, \\ z & \text{if } a \leq z \leq b, \\ b & \text{if } z > b. \end{cases}$$

Truncated procedures may be useful in a number of circumstances. For example, if the functions in the recursive equation are defined only for certain values of the parameter, then the procedure should produce points only from this set. Truncations may also be useful when certain standard assumptions, e.g., conditions on the growth rate of the relevant functions are not satisfied. Truncations may also help to make an efficient use of auxiliary information concerning the value of the unknown parameter. For example, we might have auxiliary information about the parameters, e.g. a set, possibly time dependent, that contains the value of the unknown parameter. Also, sometimes a consistent but not necessarily efficient auxiliary estimator $\tilde{\theta}_t$ is available having a rate d_t. Then to obtain asymptotically efficient estimator, one can construct a procedure with shrinking bounds by truncating the recursive procedure in a neighbourhood of θ with $[\alpha_t, \beta_t] = [\tilde{\theta}_t - \delta_t, \tilde{\theta}_t + \delta_t]$, where $\delta_t \to 0$.

Note that the idea of truncations is not new and goes back to Khasminskii and Nevelson (1972) and Fabian (1978) (see also Chen and Zhu (1986), Chen et al.(1987), Andradóttir (1995), Sharia (1997), Tadic (1997,1998), Lelong (2008). A comprehensive bibliography and some comparisons can be found in Sharia (2014)).

In order to study these procedures in an unified manner, Sharia (2014) introduced a SA of the following form

$$Z_t = \Phi_{U_t}\Big(Z_{t-1} + \gamma_t(Z_{t-1})\big[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})\big]\Big), \quad t = 1, 2, \ldots$$

where $Z_0 \in \mathbb{R}^m$ is some starting value, $R_t(z)$ is a predictable process with the property that $R_t(z^0) = 0$ for all t's, $\gamma_t(z)$ is a matrix-valued predictable step-size sequence, and $U_t \subset \mathbb{R}^m$ is a random sequence of truncation sets (see Section 2 for details). These SA procedures have the following main characteristics: (1) inhomogeneous random functions R_t; (2) state dependent matrix valued random step-sizes; (3) truncations with random and moving (shrinking or expanding) bounds. The main motivation for these comes from parametric statistical applications: (1) is needed for recursive parameter estimation procedures for non i.i.d. models; (2) is required to guarantee asymptotic optimality and efficiency of statistical estimation; (3) is needed for various different adaptive truncations, in particular, for the ones arising by auxiliary estimators.

Convergence of the above class of procedures is studied in Sharia (2014). In this paper we present new results on rate of convergence. Furthermore, we present a convergence result which generalises the corresponding result in Sharia (2014) by considering time dependent random Lyapunov type functions (see Lemma 3.1). This generalisation turns out to be quite useful as it can be used to derive convergence results of the recursive parameter estimators in time series models. Some of the conditions in the main statements are difficult to interpret. Therefore, we discuss these conditions in explanatory remarks and corollaries. The corollaries are presented in such a way that each subsequent statement imposes conditions that are more restrictive than the previous one. We discuss the case of the classical SA and demonstrate that conditions introduced in this paper are minimal in the sense that they do not impose any additional restrictions when applied to the classical case. We also compare our set of conditions to that of Kushner-Clark's setting (see Remark 4.4). Furthermore, the paper contains new results even for the classical SA. In particular, truncations with moving bounds give a possibility to use SA in the cases when the standard conditions on the function R do not hold. Also, an interesting link between the rate of the step-size sequence and the rate of convergence of the SA process is given in the classical case (see corollary 4.7 and Remark 4.8). This observation might not surprise experts working in this field, but we failed to find it in a written form in the existing literature.

2 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 $t=1,2,\dots,$ we have $(\mathcal{B}(\mathbb{R}^m)\times\mathcal{F})$-measurable functions

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

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 $E\left\{\varepsilon_t(z)\mid\mathcal{F}_{t-1}\right\}=0$. We also assume that

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

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

Suppose that $h=h(z)$ is a real valued function of $z\in\mathbb{R}^m$. Denote by $h'(z)$ the row-vector of partial derivatives of h with respect to the components of z, that is, $h'(z) = \left(\frac{\partial}{\partial z_1}h(z), \dots, \frac{\partial}{\partial z_m}h(z)\right)$. Also, we denote by $h''(z)$ the matrix of second partial derivatives. The $m \times m$ identity matrix is denoted by $\mathbf{I}$. Denote by $[a]^+$ and $[a]^-$ the positive and negative parts of $a \in \mathbb{R}$, i.e. $[a]^+ = \max(a, 0)$ and $[a]^- = \min(a, 0)$.

Let $U \subset \mathbb{R}^m$ is a closed convex set and define a truncation operator as a function $\Phi_U(z) : \mathbb{R}^m \longrightarrow \mathbb{R}^m$, such that

$$\Phi_U(z) = \begin{cases} z & \text{if } z \in U \\ z^* & \text{if } z \notin U, \end{cases}$$

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 $U_t = U_t(\omega)$ $(t = 1, 2, \dots)$ from $\mathbb{R}^m$ is <u>**admissible**</u> 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 $\Phi_{U_t}(z)$ 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

$$Z_t = \Phi_{U_t}\Big(Z_{t-1} + \gamma_t(Z_{t-1})\Psi_t(Z_{t-1})\Big), \quad t = 1, 2, \dots \tag{2.1}$$

where U_t is admissible for z^0,

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

and $R_t(z)$, $\varepsilon_t(z)$, $\gamma_t(z)$ are random fields defined above. Everywhere in this work, we assume that

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

and

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

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

Remark 2.1 Condition (2.2) ensures that $\varepsilon_t(Z_{t-1})$ 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 Kallenberg (2002)).

We say that a random field

$$V_t(z) = V_t(z,\omega) : \mathbb{R}^m \times \Omega \longrightarrow \mathbb{R} \qquad (t = 1, 2, ...)$$

is a **Lyapunov random field** if

- $V_t(z)$ is a predictable process for each $z \in \mathbb{R}^m$;
- for each t and almost all ω, $V_t(z)$ is a non-negative function with continuous and bounded partial second derivatives.

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 a 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 $t_0(\omega) < \infty$.*
- *All conditional expectations are calculated as integrals w.r.t. corresponding regular conditional probability measures.*
- *The* $\inf_{z\in U} h(z)$ *of a real valued function $h(z)$ is 1 whenever $U = \emptyset$.*

3 Convergence and rate of convergence

We start this section with a convergence lemma, which uses a concept of a Lyapunov random field (see Section 2). The proof of this lemma is very similar to that of presented in Sharia (2014). However, the dynamically changing Lyapunov functions make it possible to apply this result to derive the rate of convergence of the SA procedures. Also, this result turns out to be very useful to derive convergence of the recursive parameter estimations in time series models.

Lemma 3.1 *Suppose that Z_t is a process defined by* (2.1). *Let $V_t(u)$ be a Lyapunov random field. Denote $\Delta_t = Z_t - z^0$, $\Delta V_t(u) = V_t(u) - V_{t-1}(u)$, and assume that*

(V1)

$$V_t(\Delta_t) \leq V_t\Big(\Delta_{t-1} + \gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})]\Big)$$

eventually;

(V2)

$$\sum_{t=1}^{\infty}[1 + V_{t-1}(\Delta_{t-1})]^{-1}[\mathcal{K}_t(\Delta_{t-1})]^+ < \infty, \qquad P\text{-}a.s.,$$

where

$$\mathcal{K}_t(u) = \Delta V_t(u) + V_t'(u)\gamma_t(z^0+u)R_t(z^0+u) + \eta_t(z^0+u)$$

and

$$\eta_t(v) = \frac{1}{2}\sup_z E\left\{\Big[R_t(v)+\varepsilon_t(v)\Big]^T \gamma_t^T(v)V_t''(z)\gamma_t(v)\Big[R_t(v)+\varepsilon_t(v)\Big]\Big|\mathcal{F}_{t-1}\right\}.$$

Then $V_t(\Delta_t)$ converges (P-a.s.) to a finite limit for any initial value Z_0.

Furthermore, if there exists a set $A \in \mathcal{F}$ with $P(A) > 0$ such that for each $\epsilon \in (0,1)$

(V3)

$$\sum_{t=1}^{\infty} \inf_{\substack{\epsilon \le V_t(u) \le 1/\epsilon \\ z^0+u \in U_{t-1}}} [\mathcal{K}_t(u)]^- = \infty \quad \text{on } A, \tag{3.1}$$

then $V_t(\Delta_t) \longrightarrow 0$ (P-a.s.) for any initial value Z_0.

Proof. The proof is similar to that of Theorem 2.2 and 2.4 in Sharia (2014). Rewrite (2.1) in the form

$$\Delta_t = \Delta_{t-1} + \gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})].$$

By (V1), using the Taylor expansion, we have

$$\begin{aligned} V_t(\Delta_t) \le\ & V_t(\Delta_{t-1}) + V_t'(\Delta_{t-1})\gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})] \\ & + \frac{1}{2}[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})]^T \gamma_t^T(Z_{t-1})V_t''(\tilde{\Delta}_{t-1})\gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})], \end{aligned}$$

where $\tilde{\Delta}_{t-1} \in \mathbb{R}^m$ is $\mathcal{F}_{t-1}$-measurable Since

$$V_t(\Delta_{t-1}) = V_{t-1}(\Delta_{t-1}) + \Delta V_t(\Delta_{t-1}),$$

using (2.2) and (2.3), we obtain

$$E\{V_t(\Delta_t)|\mathcal{F}_{t-1}\} \le V_{t-1}(\Delta_{t-1}) + \mathcal{K}_t(\Delta_{t-1}).$$

Then, using the decomposition $\mathcal{K}_t = [\mathcal{K}_t]^+ - [\mathcal{K}_t]^-$, the above can be rewritten as

$$E\{V_t(\Delta_t)|\mathcal{F}_{t-1}\} \le V_{t-1}(\Delta_{t-1})(1+B_t) + B_t - [\mathcal{K}_t(\Delta_{t-1})]^-,$$

where $B_t = (1+V_{t-1}(\Delta_{t-1}))^{-1}[\mathcal{K}_t(\Delta_{t-1})]^+$.

By $(V2)$, we have that $\sum_{t=1}^{\infty} B_t < \infty$. Now we can use Lemma 6.1 in Appendix (with $X_t = V_t(\Delta_t), \beta_{t-1} = \xi_{t-1} = B_t$ and $\zeta_t = [\mathcal{K}_t(\Delta_{t-1})]^-$) to deduce that the processes $V_t(\Delta_t)$ and

$$Y_t = \sum_{s=1}^{t} [\mathcal{K}_s(\Delta_{s-1})]^-$$

converge to some finite limits. Therefore, it follows that $V_t(\Delta_t) \to r \geq 0$.

To prove the second assertion, suppose that $r > 0$. Then there exist $\epsilon > 0$ such that $\epsilon \leq V_t(\Delta_t) \leq 1/\epsilon$ eventually. By (3.1), this would imply that for some t_0,

$$\sum_{s=t_0}^{\infty} [\mathcal{K}_s(\Delta_{s-1})]^- \geq \sum_{s=t_0}^{\infty} \inf_{\substack{\epsilon \leq V_s(u) \leq 1/\epsilon \\ z^0 + u \in U_{s-1}}} [\mathcal{K}_s(u)]^- = \infty$$

on the set A, which contradicts the existence of a finite limit of Y_t. Hence, $r = 0$ and $V_t(\Delta_t) \longrightarrow 0$. ■

Remark 3.2 The conditions of the above Lemma are difficult to interpret. Therefore, the rest of the section is devoted to formulate lemmas and corollaries (Lemmas 3.5 and 3.9, Corollaries 3.7, 3.12 and 3.13) containing sufficient conditions for the convergence and the rate of convergence, and remarks (Remarks 3.3, 3.4, 3.8, 3.10, 3.11 and 3.14) explaining some of the assumptions. These results are presented in such a way, that each subsequent statement imposes conditions that are more restrictive than the previous one. For example, Corollary 3.13 and Remark 3.14 contain conditions which are most restrictive than all the previous ones, but are written in the simplest possible terms.

Remark 3.3 A typical choice of $V_t(u)$ is $V_t(u) = u^T C_t u$, where $\{C_t\}$ is a predictable positive semi-definite matrix process. If C_t/a_t goes to a finite matrix with $a_t \longrightarrow \infty$, then subject to the conditions of Lemma 3.1, $a_t \|Z_t - z^0\|^2$ will tend to a finite limit implying that $Z_t \longrightarrow z^0$. This approach is adopted in Example 5.3 to derive convergence of the on-line Least Square estimator.

Remark 3.4 Consider truncation sets $U_t = S(\alpha_t, r_t)$, where S denotes a closed sphere in $\mathbb{R}^m$ with the center at $\alpha_t \in \mathbb{R}^m$ and the radius r_t. Let $z_t' = \Phi_{U_t}(z_t)$ and suppose that $z^0 \in U_t$. Let $V_t(u) = u^T C_t u$ where C_t is a positive definite matrix and denote by λ_t^{max} and λ_t^{min} the largest and smallest eigenvalues of C_t respectively. Then $(z_t' - z^0)^T C_t (z_t' - z^0) \leq (z_t - z^0)^T C_t (z_t - z^0)$ $\Big($i.e., (V1) holds with $V_t(u) = u^T C_t u\Big)$, if $\lambda_t^{max} v_t^2 \leq \lambda_t^{min} r_t^2$, where $v_t = \|\alpha_t - z^0\|$. (See Proposition 6.2 in Appendix for details.) In particular, if C_t is a scalar matrix, condition (V1) automatically holds.

Lemma 3.5 *Suppose that all the conditions of Lemma 3.1 hold and*

(L) *for any $M > 0$, there exist some $\delta = \delta(\omega) > 0$ such that*

$$\inf_{\|u\| \geq M} V_t(u) > \delta \qquad \textit{eventually.}$$

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

Proof. From Lemma 3.1, we have $V_t(\Delta_t) \longrightarrow 0$ (a.s.). Now, $\Delta_t \longrightarrow 0$ follows from (L) by contradiction. Indeed, suppose that $\Delta_t \not\longrightarrow 0$ on a set, say B of positive probability. Then, for any fixed ω from this set, there would exist a sequence $t_k \longrightarrow \infty$ such that $\|\Delta_{t_k}\| \geq \epsilon$ for some $\epsilon > 0$, and (3.5) would imply that $V_{t_k}(\Delta_{t_k}) > \delta > 0$ for large k-s, which contradicts the P-a.s. convergence $V_t(\Delta_t) \longrightarrow 0$. ■

Remark 3.6 The following corollary contains simple sufficient conditions for convergence. The poof of this corollary does not require dynamically changing Lyapunov functions and can be obtained from a less general version of Lemma 3.1 presented in Sharia (2014). We decided to present this corollary for the sake of completeness, noting that the proof, as well as a number of different sets of sufficient conditions, can be found in Sharia (2014).

Corollary 3.7 *Suppose that Z_t is a process defined by* (2.1)*, U_t are admissible truncations for z^0 and*

(D1) *for large t's*

$$(z - z^0)^T R_t(z) \leq 0 \quad \textit{if} \quad z \in U_{t-1};$$

(D2) *there exists a predictable process $r_t > 0$ such that*

$$\sup_{z \in U_{t-1}} \frac{E\left\{\|R_t(z) + \varepsilon_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}}{1 + \|z - z^o\|^2} \leq r_t$$

eventually, and

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

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

Furthermore, if

(D3) *for each $\epsilon \in (0,1)$, there exists a predictable process $\nu_t > 0$ such that*

$$\inf_{\substack{\epsilon \leq \|z - z^o\| \leq 1/\epsilon \\ z \in U_{t-1}}} -(z - z^0)^T R_t(z) > \nu_t$$

eventually, where

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

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

Proof. See Remark 3.6 above.

Remark 3.8 The rest of this section is concerned with the derivation of sufficient conditions to establish rate of convergence. In most applications, checking conditions of Lemma 3.9 and Corollary 3.12 below is difficult without establishing the convergence of Z_t first. Therefore, although formally not required, we can assume that $Z_t \longrightarrow z^0$ convergence has already been established (using the lemmas and corollaries above or otherwise). Under this assumption, conditions for the rate of convergence below can be regarded as local in z^0, that is, they can be derived using certain continuity and differentiability assumptions of the corresponding functions at point z^0 (see examples in Section 5).

Lemma 3.9 *Suppose that Z_t is a process defined by (2.1). Let $\{C_t\}$ be a predictable positive definite $m \times m$ matrix process, and λ_t^{max} and λ_t^{min} be the largest and the smallest eigenvalues of C_t respectively. Denote $\Delta_t = Z_t - z^0$. Suppose also that (V1) of Lemma 3.1 holds and*

(R1) *there exists a predictable non-negative scalar process $\mathcal{P}_t$ such that*

$$\frac{2\Delta_{t-1}^T C_t \gamma_t(z^0 + \Delta_{t-1}) R_t(z^0 + \Delta_{t-1})}{\lambda_t^{max}} + \mathcal{P}_t \leq -\rho_t \|\Delta_{t-1}\|^2,$$

eventually, where ρ_t is a predictable non-negative scalar process satisfying

$$\sum_{t=1}^{\infty} \left[\frac{\lambda_t^{max} - \lambda_{t-1}^{min}}{\lambda_{t-1}^{min}} - \frac{\lambda_t^{max}}{\lambda_{t-1}^{min}} \rho_t \right]^+ < \infty;$$

(R2)

$$\sum_{t=1}^{\infty} \frac{\lambda_t^{max} \left[E\left\{ \left\| \gamma_t(z^0+\Delta_{t-1}) \Big[R_t(z^0+\Delta_{t-1}) + \varepsilon_t(z^0+\Delta_{t-1}) \Big] \right\|^2 \mid \mathcal{F}_{t-1} \right\} - \mathcal{P}_t \right]^+}{1+\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2} < \infty.$$

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

Proof. Let us check the conditions of Lemma 3.1 with $V_t(u) = u^T C_t u$. Condition (V1) is satisfied automatically.

Denote $R_t = R_t(z^0+\Delta_{t-1})$, $\gamma_t = \gamma_t(z^0+\Delta_{t-1})$ and $\varepsilon_t = \varepsilon_t(z^0+\Delta_{t-1})$. Since $V_t'(u) = 2u^T C_t$ and $V_t''(u) = 2C_t$, we have

$$\mathcal{K}_t(\Delta_{t-1}) = \Delta V_t(\Delta_{t-1}) + 2\Delta_{t-1}^T C_t \gamma_t R_t + E\left\{ [\gamma_t(R_t+\varepsilon_t)]^T C_t \gamma_t (R_t+\varepsilon_t) \mid \mathcal{F}_{t-1} \right\}$$

Since C_t is positive definite, $\lambda_t^{min}\|u\|^2 \le u^T C_t u \le \lambda_t^{max}\|u\|^2$ for any $u \in \mathbb{R}^m$. Therefore

$$\Delta V_t(\Delta_{t-1}) \le (\lambda_t^{max} - \lambda_{t-1}^{min})\|\Delta_{t-1}\|^2.$$

Denote

$$\tilde{\mathcal{P}}_t = \lambda_t^{max}(\mathcal{D}_t - \mathcal{P}_t)$$

where

$$\mathcal{D}_t = E\left\{ \|\gamma_t(R_t+\varepsilon_t)\|^2 \mid \mathcal{F}_{t-1} \right\}.$$

Then

$$\begin{aligned} \mathcal{K}_t(\Delta_{t-1}) &\le (\lambda_t^{max} - \lambda_{t-1}^{min})\|\Delta_{t-1}\|^2 + 2\Delta_{t-1}^T C_t \gamma_t R_t + \lambda_t^{max}\mathcal{D}_t \\ &= (\lambda_t^{max} - \lambda_{t-1}^{min})\|\Delta_{t-1}\|^2 + 2\Delta_{t-1}^T C_t \gamma_t R_t + \lambda_t^{max}\mathcal{P}_t + \tilde{\mathcal{P}}_t\ . \end{aligned}$$

By (R1), we have

$$2\Delta_{t-1}^T C_t \gamma_t R_t \le -\lambda_t^{max}(\rho_t\|\Delta_{t-1}\|^2 + \mathcal{P}_t).$$

Therefore,

$$\begin{aligned} \mathcal{K}_t(\Delta_{t-1}) &\le (\lambda_t^{max} - \lambda_{t-1}^{min})\|\Delta_{t-1}\|^2 - \lambda_t^{max}(\rho_t\|\Delta_{t-1}\|^2 + \mathcal{P}_t) + \lambda_t^{max}\mathcal{P}_t + \tilde{\mathcal{P}}_t \\ &\le (\lambda_t^{max} - \lambda_{t-1}^{min} - \lambda_t^{max}\rho_t)\|\Delta_{t-1}\|^2 + \tilde{\mathcal{P}}_t = r_t\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2 + \tilde{\mathcal{P}}_t, \end{aligned}$$

where

$$r_t = (\lambda_t^{max} - \lambda_{t-1}^{min} - \lambda_t^{max}\rho_t)/\lambda_{t-1}^{min}.$$

Since $\lambda_{t-1}^{min} \geq 0$, using the inequality $[a+b]^+ \leq [a]^+ + [b]^+$, we have

$$[\mathcal{K}_t(\Delta_{t-1})]^+ \leq \lambda_{t-1}^{min} \|\Delta_{t-1}\|^2 [r_t]^+ + [\tilde{\mathcal{P}}_t]^+.$$

Also, since $V_{t-1}(\Delta_{t-1}) = \Delta_{t-1}^T C_{t-1} \Delta_{t-1} \geq \lambda_{t-1}^{min} \|\Delta_{t-1}\|^2$,

$$\begin{aligned}\frac{[\mathcal{K}_t(\Delta_{t-1})]^+}{1+V_{t-1}(\Delta_{t-1})} &\leq \frac{[\mathcal{K}_t(\Delta_{t-1})]^+}{1+\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2} \leq \frac{\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2 [r_t]^+}{1+\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2} + \frac{[\tilde{\mathcal{P}}_t]^+}{1+\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2} \\ &\leq [r_t]^+ + \frac{[\tilde{\mathcal{P}}_t]^+}{1+\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2}.\end{aligned}$$

By (R2), $\sum_{t=1}^{\infty} [\tilde{\mathcal{P}}_t]^+ / (1 + \lambda_{t-1}^{min} \|\Delta_{t-1}\|^2) < \infty$ and according to (R1)

$$\sum_{t=1}^{\infty} [r_t]^+ = \sum_{t=1}^{\infty} \left[\frac{\lambda_t^{max} - \lambda_{t-1}^{min}}{\lambda_{t-1}^{min}} - \frac{\lambda_t^{max}}{\lambda_{t-1}^{min}} \rho_t \right]^+ < \infty.$$

Thus,

$$\sum_{t-1}^{\infty} \frac{[\mathcal{K}_t(\Delta_{t-1})]^+}{1+V_{t-1}(\Delta_{t-1})} < \infty,$$

implying that Condition (V2) of Lemma 3.1 holds. Thus, $(Z_t - z^0)^T C_t (Z_t - z^0)$ converges to a finite limit almost surely. ■

Remark 3.10 The choice $\mathcal{P}_t = 0$ means that (R2) becomes more restrictive imposing stronger probabilistic restrictions on the model. Now, if $\Delta_{t-1}^T C_t \gamma_t (z^0 + \Delta_{t-1}) R_t (z^0 + \Delta_{t-1})$ is eventually negative with a large absolute value, then it is possible to introduce a non-zero $\mathcal{P}_t$ without strengthening condition (R1). One possibility might be $\mathcal{P}_t = \|\gamma_t R_t\|^2$. In that case, since γ_t and R_t are predictable processes, and sequence ε_t is a martingale-difference,

$$E\{\|\gamma_t(R_t + \varepsilon_t)\|^2 | \mathcal{F}_{t-1}\} = \|\gamma_t R_t\|^2 + E\{\|\gamma_t \varepsilon_t\|^2 | \mathcal{F}_{t-1}\}.$$

Then condition (R2) can be rewritten as

$$\sum_{t=1}^{\infty} \lambda_t^{max} E\{\|\gamma_t(z^0 + \Delta_{t-1}) \varepsilon_t (z^0 + \Delta_{t-1})\|^2 | \mathcal{F}_{t-1}\} < \infty.$$

Remark 3.11 The next corollary is a special case of Lemma 3.9 when the step-size sequence is a sequence of scalar matrices, i.e. $\gamma_t(Z_{t-1}) = a_t^{-1} \mathbf{I}$, where a_t is non-decreasing and positive.

Corollary 3.12 *Let Z_t be a process defined by* (2.1). *Suppose that $a_t > 0$ is a non-decreasing sequence and*

(W1)

$$\Delta_{t-1}^T R_t(Z_{t-1}) \le -\frac{1}{2}\Delta a_t \|\Delta_{t-1}\|^2$$

eventually;

(W2) *there exist $0 < \delta \le 1$ such that,*

$$\sum_{t=1}^{\infty} a_t^{\delta-2} E\left\{\|(R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1}))\|^2 \mid \mathcal{F}_{t-1}\right\} < \infty.$$

Then $a_t^\delta \|Z_t - z^0\|^2$ converges to a finite limit (P-a.s.).

Proof. Consider Lemma 3.9 with $\gamma_t = \gamma_t(z) = a_t^{-1}\mathbf{I}$, $C_t = a_t^\delta \mathbf{I}$, $\mathcal{P}_t = 0$ and $\rho_t = \Delta a_t / a_t$. To check (R2), denote the infinite sum in (R2) by Q, then

$$\begin{aligned} Q &\le \sum_{t=1}^{\infty} \lambda_t^{max}\left[E\left\{\left\|\gamma_t\Big[R_t(z^0+\Delta_{t-1}) + \varepsilon_t(z^0+\Delta_{t-1})\Big]\right\|^2 \mid \mathcal{F}_{t-1}\right\} - \mathcal{P}_t\right]^+ \\ &\le \sum_{t=1}^{\infty} \lambda_t^{max}\|\gamma_t\|^2 E\left\{\|(R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1}))\|^2 \mid \mathcal{F}_{t-1}\right\}. \end{aligned}$$

Now, since $\lambda_t^{min} = \lambda_t^{max} = a_t^\delta$ and $\|\gamma_t\|^2 = a_t^{-2}$, condition (W2) leads to (R2).

Since $\rho_t = \Delta a_t / a_t < 1$ and $(a_t/a_{t-1})^\delta \le a_t/a_{t-1}$,

$$\begin{aligned} \sum_{t=1}^{\infty}\left[\frac{\lambda_t^{max} - \lambda_{t-1}^{min}}{\lambda_{t-1}^{min}} - \frac{\lambda_t^{max}}{\lambda_{t-1}^{min}}\rho_t\right]^+ &= \sum_{t=1}^{\infty}\left[\frac{a_t^\delta - a_{t-1}^\delta}{a_{t-1}^\delta} - \frac{a_t^\delta}{a_{t-1}^\delta}\rho_t\right]^+ \\ &= \sum_{t=1}^{\infty}\left[(1-\rho_t)\frac{a_t^\delta}{a_{t-1}^\delta} - 1\right]^+ \\ &\le \sum_{t=1}^{\infty}\left[(1 - \frac{\Delta a_t}{a_t})\frac{a_t}{a_{t-1}} - 1\right]^+ = 0\,. \end{aligned}$$

Therefore, (W1) leads to (R1). According to Remark 3.4, condition (V1) holds since $V_t(u) = a_t^\delta \|u\|^2$. Thus, all the conditions of Lemma 3.9 hold and $a_t^\delta \|Z_t - z^0\|^2$ converges to a finite limit (P-a.s.). ■

Corollary 3.13 *Let Z_t be a process defined by (2.1) where $z^0 \in \mathbb{R}$, $\gamma_t(Z_{t-1}) = 1/t$ and the truncation sequence U_t is admissible. Suppose that $Z_t \longrightarrow z^0$ and*

(Y1) *$R_t'(z^0) \leq -1/2$ eventually;*

(Y2) *$R_t(z)$ and $\sigma_t^2(z) = E(\varepsilon_t^2(z)|\mathcal{F}_{t-1})$ are locally uniformly bounded at z^0 w.r.t. t; that is, there exists a constant K such that $|R_t(\xi_t)| \leq K$ and $|\sigma_t^2(\xi_t)| \leq K$ eventually, for any $\xi_t \longrightarrow z^0$.*

Then $t^\delta(Z_t - z^0)^2$ converges to a finite limit (P-a.s.), for any $\delta < 1$.

Proof. Consider Corollary 3.12 with $a_t = t$. In the one-dimensional case, condition (W1) can be rewritten as

$$\frac{R_t(z^0 + \Delta_{t-1})}{\Delta_{t-1}} \leq -\frac{1}{2}.$$

Condition (W1) now follows from (Y1).

Since $E\{\varepsilon_t(z)|\mathcal{F}_{t-1}\} = 0$, using (Y2) we have for any $\delta < 1$,

$$\begin{aligned} &\sum_{t=1}^{\infty} t^{\delta-2} E\left\{(R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1}))^2 \mid \mathcal{F}_{t-1}\right\} \\ = \; &\sum_{t=1}^{\infty} t^{\delta-2} R_t^2(Z_{t-1}) + \sum_{t=1}^{\infty} t^{\delta-2} E\left\{\varepsilon_t^2(Z_{t-1}) \mid \mathcal{F}_{t-1}\right\} < \infty. \end{aligned}$$

Thus, condition (W2) holds. Therefore, $t^\delta(Z_t - z^0)^2$ converges to a finite limit (P-a.s.), for any $\delta < 1$. ■

Remark 3.14 Corollary 3.13 gives simple but more restrictive sufficient conditions to derive the rate of convergence in one-dimensional cases. It is easy to see that all conditions of Corollary 3.13 trivially hold, if e.g., ε_t are state independent i.i.d. random variables with a finite second moment, $R_t(z) = R(z)$, and $R'(z^0) \leq -1/2$.

4 Classical problem stochastic approximation

Consider the classical problem of stochastic approximation to find a root z^0 of the equation $R(z^0) = 0$. Let us take a step-size sequence $\gamma_t = a_t^{-1}\mathbf{I}$, where $a_t \longrightarrow \infty$ is a predictable scalar process, and consider the procedure

$$Z_t = \Phi_{U_t}\Big(Z_{t-1} + a_t^{-1}[R(Z_{t-1}) + \varepsilon_t(Z_{t-1})]\Big). \tag{4.1}$$

Corollary 4.1 *Suppose that Z_t is a process defined by* (4.1)*, truncation sequence U_t is admissible, and*

(H1)
$$(z - z^0)^T R(z) \le 0$$
for any $z \in \mathbb{R}^m$ with the property that $z \in U_t$ eventually;

(H2) there exists a predictable process r_t such that
$$\sup_{z \in U_{t-1}} \|R(z)\| \le r_t \quad \textit{where} \quad \sum_{t=1}^{\infty} a_t^{-2} r_t < \infty;$$

(H3) there exists a predictable process e_t such that
$$\sup_{z \in U_{t-1}} \frac{E\{\|\varepsilon_t(z)\|^2 | \mathcal{F}_{t-1}\}}{1 + \|z - z^0\|^2} \le e_t$$
eventually, where
$$\sum_{t=1}^{\infty} e_t a_t^{-2} < \infty \;\; P\textit{-a.s.}.$$

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

Furthermore, suppose that

(H4) $R(z)$ is continuous at z^0 and $(z - z^0)^T R(z) < 0$ for all z with the property that $z \in U_t \backslash \{z^0\}$ eventually;

(H5)
$$\sum_{t=1}^{\infty} a_t^{-1} = \infty.$$

Then $Z_t \longrightarrow z^0$ (P-a.s.).

Proof. Consider Corollary 3.7 with $R_t = R$. Condition (D1) trivially holds. Since $E\left\{\varepsilon_t(u) \mid \mathcal{F}_{t-1}\right\} = 0$, we have
$$E\left\{\|R(z) + \varepsilon(z)\|^2 \mid \mathcal{F}_{t-1}\right\} = \|R(z)\|^2 + E\left\{\|\varepsilon_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}.$$
Now condition (D2) holds with $p_t = r_t + e_t$.

By (H4), there exists a constant $\nu > 0$ such that for each $\epsilon \in (0,1)$

$$\inf_{\substack{\varepsilon \leq \|z-z^o\| \leq 1/\varepsilon \\ z \in U_{t-1}}} -(z-z^0)^T R(u) > \nu$$

eventually and by (H5) $\sum_{t=1}^{\infty} \nu a_t^{-1} = \nu \sum_{t=1}^{\infty} a_t^{-1} = \infty$. This implies that (D3) also holds. Therefore, by Corollary 3.7, $Z_t \longrightarrow z^0$ almost surely. ■

Remark 4.2 Suppose that $\varepsilon_t = \varepsilon_t(z)$ 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 (H3) holds if

$$\sum_{t=1}^{\infty} \sigma_t^2 a_t^{-2} < \infty, \qquad P\text{-a.s.}. \tag{4.2}$$

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 4.3 (a) If the truncation sets are uniformly bounded, then some of the conditions above can be weakened considerably. For example, condition (H2) in Corollary 4.1 will automatically hold given that $\sum_{t=1}^{\infty} a_t^{-2} < \infty$.
(b) Also if it is only required that Z_t converges to any finite limit, the step-size sequence a_t can go to infinity at any rate as long as $\sum_{t=1}^{\infty} a_t^{-2} < \infty$. However, in order to have $Z_t \longrightarrow z^0$, one must ensure that a_t does not increase too fast. Also, the variances of the error terms can go to infinity as t tends to infinity, as long as the sum in (H3) is bounded.

Remark 4.4 To compare the above result to that of Kushner-Clark's setting, let us assume boundedness of Z_t. Then there exists a compact set U such that $Z_t \in U$. Without lost of generality, we can assume that $z^0 \in U$. Then Z_t in Corollary 4.1 can be assumed to be generated using the truncations on $U_t \cap U$. Let us assume that $\sum_{s=1}^{\infty} a_t^{-2} < \infty$. Then, condition (H2) will hold if, e.g., $R(z)$ is a continuous function. Also, in this case, given that the error terms $\varepsilon_t(z)$ are continuous in z with some uniformity w.r.t. t, they will in fact behave in the same way as state independent error terms. Therefore, a condition of the type (4.2) given in Remark 4.2 will be sufficient for (H3).

Corollary 4.5 *Suppose that Z_t, defined by* (4.1)*, converges to z^0 (P-a.s.) and truncation sequence U_t is admissible. Suppose also that*

(B1)

$$u^T R(z^0+u) \le -\frac{1}{2}\|u\|^2 \quad \textit{for small } u\textit{'s;}$$

(B2) $a_t > 0$ *is non-decreasing with*

$$\sum_{t=1}^{\infty}\left[\frac{\Delta a_t - 1}{a_{t-1}}\right]^+ < \infty;$$

(B3) there exist $\delta \in (0,1)$ *such that*

$$\sum_{t=1}^{\infty} a_t^{\delta-2}\|R(z^0+v_t)\|^2 < \infty \qquad \textit{and} \qquad \sum_{t=1}^{\infty} a_t^{\delta-2} E\{\|\varepsilon_t(z^0+v_t)\|^2|\mathcal{F}_{t-1}\} < \infty,$$

where $v_t \in U_t$ *is any predictable process with the property* $v_t \longrightarrow 0$.

Then $a_t^\delta\|Z_t - z^0\|^2$ *converges (P-a.s.) to a finite limit.*

Proof. Let us check that conditions of Lemma 3.9 hold with $R_t = R$, $\rho_t = a_t^{-1}$, $\mathcal{P}_t = 0$ and $C_t = a_t^\delta \mathbf{I}$. We have $\lambda_t^{max} = \lambda_t^{min} = a_t^\delta$ by (B2), and

$$\begin{aligned}
&\sum_{t=1}^{\infty}\left[\frac{\lambda_t^{max}-\lambda_{t-1}^{min}}{\lambda_{t-1}^{min}} - \frac{\lambda_t^{max}}{\lambda_{t-1}^{min}}\rho_t\right]^+ = \sum_{t=1}^{\infty}\left[\frac{a_t^\delta - a_{t-1}^\delta}{a_{t-1}^\delta} - \frac{a_t^\delta}{a_{t-1}^\delta a_t}\right]^+ \\
= \; & \sum_{t=1}^{\infty}\left[\left(\frac{a_t}{a_{t-1}}\right)^\delta (1-a_t^{-1}) - 1\right]^+ \le \sum_{t=1}^{\infty}\left[\frac{a_t}{a_{t-1}}(1-a_t^{-1}) - 1\right]^+ + C \\
= \; & \sum_{t=1}^{\infty}\left[\frac{\Delta a_t - 1}{a_{t-1}}\right]^+ + C < \infty
\end{aligned}$$

for some constant C. So (B1) leads to (R1). Also since $Z_t \longrightarrow z^0$,

$$\begin{aligned}
&\sum_{t=1}^{\infty}\frac{\lambda_t^{max}\left[E\left\{\|\gamma_t(R_t+\varepsilon_t)\|^2 \mid \mathcal{F}_{t-1}\right\} - \mathcal{P}_t\right]^+}{1+\lambda_{t-1}^{min}\|\Delta_{t-1}\|^2} \\
\le \; & \sum_{t=1}^{\infty}\lambda_t^{max}\left[E\left\{\|\gamma_t(R_t+\varepsilon_t)\|^2 \mid \mathcal{F}_{t-1}\right\} - \mathcal{P}_t\right]^+ \\
= \; & \sum_{t=1}^{\infty} a_t^\delta E\left\{\|a_t^{-1}(R_t+\varepsilon_t)\|^2 \mid \mathcal{F}_{t-1}\right\} \\
\le \; & \sum_{t=1}^{\infty} a_t^{\delta-2}\|R(Z_{t-1})\|^2 + \sum_{t=1}^{\infty} a_t^{\delta-2} E\{\|\varepsilon_t(Z_{t-1})\|^2|\mathcal{F}_{t-1}\},
\end{aligned}$$

condition (R2) follows from (B3). Therefore by Lemma 3.9, $(Z_t - z^0)^T C_t (Z_t - z^0) = a_t^\delta \|Z_t - z^0\| \longrightarrow 0$ (P-a.s.). ■

Remark 4.6 It follows from Proposition 6.3 in Appendix that if $a_t = t^\epsilon$ with $\epsilon > 1$, then (B2) doesn't hold. However, condition (B2) holds if $a_t = t^\epsilon$ for all $\epsilon \le 1$. Indeed,

$$\begin{aligned} \sum_{t=1}^{\infty} \left[\frac{\Delta a_t - 1}{a_{t-1}} \right]^+ &= \sum_{t=1}^{\infty} \left[\left(\frac{t}{t-1} \right)^\epsilon - 1 - \frac{1}{(t-1)^\epsilon} \right]^+ \\ &\le \sum_{t=1}^{\infty} \left[\frac{t}{t-1} - 1 - \frac{1}{t-1} \right]^+ = 0. \end{aligned}$$

Corollary 4.7 *Suppose that $Z_t \longrightarrow z^0$, where Z_t is defined by (4.1) with $a_t = t^\epsilon$ where $\epsilon \in (1/2, 1]$, and (B1) in Corollary 4.5 holds. Suppose also that R is continuous at z^0 and there exists $0 < \delta < 2 - 1/\epsilon$ such that*

(BB)

$$\sum_{t=1}^{\infty} \frac{1}{t^{(2-\delta)\epsilon}} E\{\|\varepsilon_t(z^0 + v_t)\|^2 | \mathcal{F}_{t-1}\} < \infty.$$

where $v_t \in U_t$ is any predictable process with the property $v_t \longrightarrow 0$.

Then $t^\delta \|Z_t - z^0\|^2$ converges to a finite limit (P-a.s.).

Proof. Let us check conditions of Corollary 4.5 with $a_t = t^\epsilon$ where $\epsilon \in (1/2, 1]$. Condition (B2) is satisfied (See Remark 4.6). Since R is continuous at z^0 and $Z_t \longrightarrow z^0$, it follows that $R(z^0 + v_t)$ in (B3) is bounded. Also, $a_t^{\delta-2} = t^{(\delta-2)\varepsilon}$ and since $(\delta - 2)\epsilon < -1$, it follows that the first part of (B3) holds. The second part is a consequence of (BB). The result is now immediate from Corollary 4.5. ■

Remark 4.8 Suppose that $a_t = t^\varepsilon$ with $\varepsilon \in (1/2, 1)$ and $\sup_t E\{\|\varepsilon_t(z)\|^2 | \mathcal{F}_{t-1}\} < \infty$ (e.g., assume that $\varepsilon_t = \varepsilon_t(z)$ are state independent and i.i.d.). Then, since $(\delta - 2)\epsilon < -1$, condition (BB) in Corollary 4.7 automatically holds for any $\delta < 2 - 1/\epsilon$. It therefore follows that the step-size sequence $a_t = t^\epsilon$, $\epsilon \in (1/2, 1)$ produces SA procedures which converge with the rate $t^{-\alpha}$ where $\alpha < 1 - \frac{1}{2\epsilon}$. For example, the step-size $a_t = t^{3/4}$ would produce the SA procedures, which converge with the rate $t^{-1/3}$.

5 Special models and examples

5.1 Finding a root of a polynomial

Let l be a positive integer and

$$R(z) = -\sum_{i=1}^{l} C_i (z - z^0)^i,$$

where $z, z^0 \in \mathbb{R}$ and C_i are real constants. Suppose that

$$(z - z^0) R(z) \le 0 \quad \text{for all} \quad z \in \mathbb{R}.$$

Note that if $l > 1$, the SA without truncations fails to satisfy the standard condition on the rate of growth at infinity. Therefore, one needs to use slowly expanding truncations to slow down the growth of R at infinity. Consider Z_t defined by (4.1) with a truncation sequence $U_t = [-u_t, u_t]$, where $u_t \longrightarrow \infty$ is a sequence of non-decreasing positive numbers. Suppose that

$$\sum_{t=1}^{\infty} u_t^{2l} \ a_t^{-2} < \infty. \tag{5.1}$$

Then, provided that the measurement errors satisfy condition (H3) of Corollary 4.1, $|Z_t - z^0|$ converges (P-a.s.) to a finite limit.

Indeed, condition (H1) of Corollary 4.1 trivially holds. For large t's,

$$\begin{aligned} \sup_{z \in [-u_{t-1}, u_{t-1}]} \|R(z)\|^2 \ &\le \sup_{z \in [-u_{t-1}, u_{t-1}]} \left[\sum_{i=1}^{l} C_i (z - z^0)^i \right]^2 \\ \le \sup_{z \in [-u_{t-1}, u_{t-1}]} \sum_{i=1}^{l} C_i^2 (z - z^0)^{2i} &\le \sum_{i=1}^{l} C_i^2 (2u_{t-1})^{2i} \le l 4^l C_l^2 u_{t-1}^{2l}, \end{aligned}$$

which, by (5.1), implies condition (H2) of Corollary 4.1.

Furthermore, if z^0 is a unique root, then provided that

$$\sum_{t=1}^{\infty} a_t^{-1} = \infty, \tag{5.2}$$

it follows from Corollary 4.1 that $Z_t \longrightarrow z^0$ (P-a.s.). One can always choose a suitable truncation sequence which satisfies (5.1) and (5.2). For example, if the degree of the polynomial is known to be l (or at most l), and $a_t = t$, then one can take $u_t = Ct^{r/2l}$, where C and r are some positive constants and $r < 1$. One can also take a truncation sequence which is independent of l, e.g., $u_t = C \log t$, where C is a positive constant.

Suppose also that

$$C_1 \geq \frac{1}{2}, \quad a_t = t^{\epsilon} \quad \text{where} \quad \epsilon \in (0, 1]$$

and condition (BB) in Corollary 4.7 holds (e.g., one can assume for simplicity that ε_t's are state independent and i.i.d.). Then $t^{\alpha}(Z_t - z^0) \xrightarrow{a.s.} 0$ for any $\alpha < 1 - 1/2\epsilon$.

Indeed, since $R'(z^0) = -C_1 \leq -1/2$, condition (B1) of Corollary 4.5 holds. Now, the above convergence is a consequence of Corollary 4.7 and Remark 4.8.

5.2 Linear procedures

Consider the recursive procedure

$$Z_t = Z_{t-1} + \gamma_t(h_t - \beta_t Z_{t-1}) \tag{5.3}$$

where γ_t is a predictable positive definite matrix process, β_t is a predictable positive semi-definite matrix process and h_t is an adapted vector process (i.e., h_t is $\mathcal{F}_t$-measurable for $t \geq 1$). If we assume that $E\{h_t|\mathcal{F}_{t-1}\} = \beta_t z^0$, we can view (5.3) as a SA procedure designed to find the common root z^0 of the linear functions

$$R_t(u) = E\{h_t - \beta_t u|\mathcal{F}_{t-1}\} = E\{h_t|\mathcal{F}_{t-1}\} - \beta_t u = \beta_t(z^0 - u)$$

which is observed with the random noise

$$\varepsilon_t(u) = h_t - \beta_t u - R_t(u) = h_t - E\{h_t|\mathcal{F}_{t-1}\} = h_t - \beta_t z^0.$$

Corollary 5.1 *Suppose that Z_t is defined by* (5.3) *with $E(h_t|\mathcal{F}_{t-1}) = \beta_t z^0$. Suppose also that a_t is a non-decreasing positive predictable process and*

(G1) $\Delta\gamma_t^{-1} - 2\beta_t + \beta_t\gamma_t\beta_t$ *is negative semi-definite eventually;*

(G2)

$$\sum_{t=1}^{\infty} a_t^{-1} E\{(h_t - \beta_t z^0)^T \gamma_t (h_t - \beta_t z^0)|\mathcal{F}_{t-1}\} < \infty.$$

Then $a_t^{-1}(Z_t - z^0)^T \gamma_t^{-1}(Z_t - z^0)$ converges to a finite limit (P-a.s.).

Proof. Let us show that conditions of Lemma 3.1 hold with $V_t(u) = a_t^{-1} u^T \gamma_t^{-1} u$. Condition (V1) trivially holds. We have $V_t'(u) = 2a_t^{-1} u^T \gamma_t^{-1}$, $V_t''(u) = 2a_t^{-1}\gamma_t^{-1}$, $R_t(z^0 + u) = -\beta_t u$ and $R_t(u) + \varepsilon_t(u) = h_t - \beta_t u$. Since $E(h_t - \beta_t z^0 | \mathcal{F}_{t-1}) = 0$, for η_t defined in (V2) we have

$$\begin{aligned}\eta_t(z^0+u) &= a_t^{-1} E\Big\{(h_t - \beta_t z^0 - \beta_t u)^T \gamma_t (h_t - \beta_t z^0 - \beta_t u)\Big| \mathcal{F}_{t-1}\Big\} \\ &= a_t^{-1} E\Big\{(h_t - \beta_t z^0)^T \gamma_t (h_t - \beta_t z^0)\Big| \mathcal{F}_{t-1}\Big\} + a_t^{-1} (\beta_t u)^T \gamma_t (\beta_t u)\ .\end{aligned}$$

Also,

$$\Delta V_t(u) = u^T [(a_t\gamma_t)^{-1} - (a_{t-1}\gamma_{t-1})^{-1}] u \le u^T (a_t\gamma_t)^{-1} u - u^T (a_t \gamma_{t-1})^{-1} u = u^T a_t^{-1} \Delta \gamma_t^{-1} u.$$

Denoting

$$\mathcal{J}_t = a_t^{-1} E\Big\{(h_t - \beta_t z^0)^T \gamma_t (h_t - \beta_t z^0)\Big| \mathcal{F}_{t-1}\Big\},$$

for $\mathcal{K}_t$ from (V2), we have

$$\begin{aligned}\mathcal{K}_t(u) &\le a_t^{-1} u^T \Delta\gamma_t^{-1} u - 2a_t^{-1} u^T \beta_t u + a_t^{-1} u^T \beta_t^T \gamma_t \beta_t u + \mathcal{J}_t \\ &= a_t^{-1} u^T (\Delta\gamma_t^{-1} - 2\beta_t + \beta_t^T \gamma_t \beta_t) u + \mathcal{J}_t\ .\end{aligned}$$

Condition (V2) is now immediate from (G1) and (G2) since

$$[1 + V_{t-1}(\Delta_{t-1})]^{-1} [\mathcal{K}_t(\Delta_{t-1})]^+ \le [\mathcal{K}_t(\Delta_{t-1})]^+ \le \mathcal{J}_t\ .$$

Thus, all the conditions of Lemma 3.1 hold which implies the required result. ■

Corollary 5.2 *Suppose that $\Delta\gamma_t^{-1} = \beta_t$, then (G1) in Corollary 5.1 holds.*

Proof. Since $\Delta\gamma_t^{-1}$ is positive semi-definite, it follows that $\Delta\gamma_t$ is negative semi-definite $\Big($see Horn and Johnson (1985) Corollary 7.7.4(a)$\Big)$. Also since

$$\begin{aligned}\Delta\gamma_t^{-1} - 2\beta_t + \beta_t\gamma_t\beta_t &= -\Delta\gamma_t^{-1} + \Delta\gamma_t^{-1}\gamma_t\Delta\gamma_t^{-1} = -\Delta\gamma_t^{-1} + \gamma_t^{-1} - 2\gamma_{t-1}^{-1} + \gamma_{t-1}^{-1}\gamma_t\gamma_{t-1}^{-1} \\ &= -\gamma_{t-1}^{-1} + \gamma_{t-1}^{-1}(\gamma_{t-1} + \Delta\gamma_t)\gamma_{t-1}^{-1} = \gamma_{t-1}^{-1}\Delta\gamma_t\gamma_{t-1}^{-1},\end{aligned}$$

it follows that (G1) holds. ■

5.3 Parameter estimation in Autoregressive models

Consider an AR(m) process

$$X_t = \theta^{(1)} X_{t-1} + \theta^{(2)} X_{t-2} + \cdots + \theta^{(m)} X_{t-m} + \xi_t = \theta^T X_{t-m}^{t-1} + \xi_t$$

where $\theta = (\theta^{(1)}, ..., \theta^{(m)})^T$, $X_{t-m}^{t-1} = (X_{t-1}, ..., X_{t-m})^T$ and ξ_t is a martingale-difference (i.e., $E\{\xi_t | \mathcal{F}_{t-1}\} = 0$). If the pdf of ξ_t w.r.t. Lebesgue's measure is $g_t(x)$, then the conditional probability density function of X_t given the past observations is

$$f_t(x, \theta | X_1^{t-1}) = f_t(x, \theta | X_{t-m}^{t-1}) = g_t(x - \theta^T X_{t-m}^{t-1})$$

and

$$\frac{f_t'^T(\theta, x | X_1^{t-1})}{f_t(\theta, x | X_1^{t-1})} = -\frac{g_t'(x - \theta^T X_{t-m}^{t-1})}{g_t(x - \theta^T X_{t-m}^{t-1})} X_{t-m}^{t-1}.$$

It is easy to see that the conditional Fisher information (1.4) is

$$I_t = \sum_{s=1}^{t} l_{gs} X_{s-m}^{s-1} (X_{s-m}^{s-1})^T \quad \text{where} \quad l_{gt} = \int_{-\infty}^{\infty} \left(\frac{g_t'(x)}{g_t(x)} \right)^2 g_t(x) dx.$$

The inverse I_t^{-1} can also be generated recursively by

$$I_t^{-1} = I_{t-1}^{-1} - l_{gt} I_{t-1}^{-1} X_{t-m}^{t-1} (1 + l_{gt} (X_{t-m}^{t-1})^T I_{t-1}^{-1} X_{t-m}^{t-1})^{-1} (X_{t-m}^{t-1})^T I_{t-1}^{-1}. \tag{5.4}$$

(Note that this can be derived either directly, or using the matrix inversion formula, sometimes referred to as the Sherman-Morrison formula.)

Thus, the on-line likelihood procedure in this case can be derived by the following recursion

$$\hat{\theta}_t = \hat{\theta}_{t-1} - I_t^{-1} X_{t-m}^{t-1} \frac{g_t'}{g_t} (X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1}) \tag{5.5}$$

where I_t^{-1} is also derived on-line using formula (5.4). In general, to include robust estimation procedures, and also to use any available auxiliary information, one can use the following class of procedures

$$\hat{\theta}_t = \Phi_{U_t} \left(\hat{\theta}_{t-1} + \gamma_t H(X_{t-m}^{t-1}) \varphi_t (X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1}) \right), \tag{5.6}$$

where $\varphi_t : \mathbb{R} \mapsto \mathbb{R}$ and $H : \mathbb{R}^m \mapsto \mathbb{R}^m$ are suitably chosen functions and γ_t is an $m \times m$ matrix valued step-size sequence.

Example 5.3 *(Recursive least squares procedures)* Recursive least squares (RLS) estimator of $\theta = (\theta^{(1)}, \dots, \theta^{(m)})^T$ is generated by the following procedure

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \hat{I}_t^{-1} X_{t-m}^{t-1} [X_t - (X_{t-m}^{t-1})^T \hat{\theta}_{t-1}], \tag{5.7}$$

$$\hat{I}_t^{-1} = \hat{I}_{t-1}^{-1} - \hat{I}_{t-1}^{-1} X_{t-m}^{t-1} [1 + (X_{t-m}^{t-1})^T \hat{I}_{t-1}^{-1} X_{t-m}^{t-1}]^{-1} (X_{t-m}^{t-1})^T \hat{I}_{t-1}^{-1}, \tag{5.8}$$

where $\hat{\theta}_0$ and a positive definite $\hat{I}_0^{-1}$ are some starting values. Note that (5.7) is a particular case of (5.6), and it also coincides with the maximum likelihood procedure (5.5) in the case when ξ_t are i.i.d. Gaussian r.v.'s.

Corollary 5.4 *Consider $\hat{\theta}_t$ defined by* (5.7) *and* (5.8). *Suppose that there exists a non-decreasing sequence $a_t > 0$ such that*

$$\sum_{t=1}^{\infty} a_t^{-1} (X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} E\{\xi_t^2 | \mathcal{F}_{t-1}\} < \infty.$$

Then $a_t^{-1} (\hat{\theta}_t - \theta)^T \hat{I}_t (\hat{\theta}_t - \theta)$ converges to a finite limit (P^θ-a.s.).

Proof. Let us check the condition of Corollary 5.1. Obviously, the matrix $\gamma_t = \hat{I}_t^{-1} = \hat{I}_0^{-1} + \sum_{s=1}^{t} X_{s-m}^{s-1} (X_{s-m}^{s-1})^T$ is positive definite and $\Delta \hat{I}_t^{-1} = \beta_t = X_{t-m}^{t-1} (X_{t-m}^{t-1})^T$ is positive semi-definite. By Corollary 5.2, condition (G1) holds. We also have

$$\sum_{t=1}^{\infty} a_t^{-1} E\{(h_t - \beta_t z^0)^T \gamma_t (h_t - \beta_t z^0) | \mathcal{F}_{t-1}\} = \sum_{t=1}^{\infty} a_t^{-1} E\{\xi_t (X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} \xi_t | \mathcal{F}_{t-1}\}$$

$$= \sum_{t=1}^{\infty} a_t^{-1} (X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} E\{\xi_t^2 | \mathcal{F}_{t-1}\} < \infty.$$

So condition (G2) holds. Hence all conditions of Corollary 5.1 hold which completes the proof. ■

Corollary 5.5 *Consider $\hat{\theta}_t$ defined by* (5.7) *and* (5.8). *Suppose that*

(P1) *there exists a non-decreasing sequence $\kappa_t \longrightarrow \infty$ such that*

$$\hat{I}_t / \kappa_t \longrightarrow G$$

where $G < \infty$ is a positive definite $m \times m$ matrix;

(P2) *there exists $\epsilon^0 \in [0, 1)$ such that*

$$E\left\{\xi_t^2 | \mathcal{F}_{t-1}\right\} \le \kappa_t^{\epsilon^0} \quad \textit{eventually.}$$

Then $\kappa_t^{1-\delta}\|\hat{\theta}_t-\theta\|^2 \longrightarrow 0$ *(P^θ-a.s.) for all* $\delta \in (\epsilon^0, 1]$.

Proof. Consider Corollary 5.4 with $a_t = \kappa_t^\delta$ for a certain $\delta \in (\epsilon^0, 1]$. By (P2), there exists t^0 such that

$$\sum_{t=t^0}^{\infty} a_t^{-1}(X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} E\{\xi_t^2|\mathcal{F}_{t-1}\} \le \sum_{t=t^0}^{\infty} \kappa_t^{\epsilon^0-\delta}(X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1}$$

eventually. Now, using (P1) and Lemma 6.4 in Appendix , the above sum converges to a finite limit implying conditions of Corollary 5.4 hold. Therefore, $(\hat{\theta}_t-\theta)^T\hat{I}_t(\hat{\theta}_t-\theta)/\kappa_t^\delta$ tends to a finite limit. Now, the assertion of the corollary follows since $\hat{I}_t/\kappa_t$ converges to a finite matrix. ■

Remark 5.6 **(a)** If X_t is a strongly stationary process, condition (P1) will trivially hold with $\kappa_t = t$. However, using the results given above, convergence can be derived without the stationarity requirement as long as $\kappa_t^{-1}\sum_{t=1}^{\infty} X_{t-m}^{t-1}(X_{t-m}^{t-1})^T$ tends to a positive define matrix.
(b) Condition (P2) demonstrates that the requirements on the innovations ξ_t are quite week. In particular, the conditional variances of the innovations do not have to be bounded w.r.t. t. For example, if $\kappa_t = t$ and the variances go to infinity not faster than t^{ε_0} (for some $0 \le \varepsilon_0 < 1$), then it follows that $t^{1-\delta}\|\hat{\theta}_t-\theta\|^2 \to 0$ for any $\delta \in (\varepsilon_0, 1)$.
(c) It follows from (a) and (b) above that in the case of a strongly stationary X_t with iid innovations, $t^{1-\delta}\|\hat{\theta}_t-\theta\|^2 \to 0$ for any $\delta > 0$ without any additional assumptions.

6 Appendix

Lemma 6.1 *Let $\mathcal{F}_0$, $\mathcal{F}_1$, … be an non-decreasing sequence of σ-algebras and X_n, β_n, ξ_n, $\zeta_n \in \mathcal{F}_n$, $n \ge 0$, be non-negative random valuables such that*

$$E(X_n|\mathcal{F}_{n-1}) \le X_{n-1}(1+\beta_{n-1}) + \xi_{n-1} - \zeta_{n-1}, \quad n \ge 1$$

eventually. Then

$$\left\{\sum_{i=1}^{\infty}\xi_{i-1} < \infty\right\} \cap \left\{\sum_{i=1}^{\infty}\beta_{i-1} < \infty\right\} \subseteq \{X \to\} \cap \left\{\sum_{i=1}^{\infty}\zeta_{i-1} < \infty\right\} \qquad P\text{-a.s.},$$

where $\{X \to\}$ denotes the set where $\lim_{n\to\infty} X_n$ exists and is finite.

Proof. The proof can be found in Robbins and Siegmund (1985). Note also that this lemma is a special case of the theorem on the convergence sets of non-negative semi-martingales (see, e.g., Lazrieva et al (1997)). ■

Proposition 6.2 *Consider a closed sphere $U = S(\alpha, r)$ in $\mathbb{R}^m$ with the center at $\alpha \in \mathbb{R}^m$ and the radius r. Let $z^0 \in U$ and $z \notin U$. Denote by z' the closest point form z to U, that is,*

$$z' = \alpha + \frac{r}{\|z - \alpha\|}(z - \alpha).$$

Suppose also that C is a positive definite matrix such that

$$\lambda_C^{max} v^2 \le \lambda_C^{min} r^2,$$

where λ_C^{max} and λ_C^{min} are the largest and smallest eigenvalues of C respectively and $v = \|\alpha - z^0\|$. Then

$$(z' - z^0)^T C(z' - z^0) \le (z - z^0)^T C(z - z^0).$$

Proof. For $u, v \in \mathbb{R}^m$, define

$$\|u\|_C = (u^T C u)^{1/2} \quad \text{and} \quad (u, v)_C = (u^T C v)^{1/2}.$$

We have

$$\begin{aligned} |(z_0 - \alpha, z' - \alpha)_C| &\le \|z_0 - \alpha\|_C \|z' - \alpha\|_C \le \sqrt{\lambda_C^{\max}}\, v \|z' - \alpha\|_C \\ &\le \sqrt{\lambda_C^{\min}}\, r \|z' - \alpha\|_C = \sqrt{\lambda_C^{\min}}\, \|z' - \alpha\| \|z' - \alpha\|_C \le \|z' - \alpha\|_C^2 \,. \end{aligned} \tag{6.1}$$

Since $z \notin U$, we have

$$z' = \alpha + \frac{r}{\|z - \alpha\|}(z - \alpha) = (1 - \delta)\alpha + \delta z,$$

where $\delta = r / \|z - \alpha\| < 1$. Then, since

$$z - z' = (1 - \delta)(z - \alpha), \quad z' - \alpha = \delta(z - \alpha), \quad z - z' = \frac{1 - \delta}{\delta}\,(z' - \alpha),$$

by (6.1),

$$\begin{aligned} (z' - z_0, z - z')_C &= (z' - \alpha, z - z')_C + (\alpha - z_0, z - z')_C \\ &= \frac{1 - \delta}{\delta}\, \|z' - \alpha\|_C^2 - \frac{1 - \delta}{\delta}\,(z_0 - \alpha, z' - \alpha)_C \ge 0. \end{aligned}$$

Therefore, since $z' - z_0 = (z - z_0) - (z - z')$, we get

$$\begin{aligned} \|z' - z_0\|_C^2 &= \|z - z_0\|_C^2 + \|z - z'\|_C^2 - 2(z - z_0, z - z')_C \\ &= \|z - z_0\|_C^2 + \|z - z'\|_C^2 - 2\|z - z'\|_C^2 - 2(z' - z_0, z - z')_C \\ &= \|z - z_0\|_C^2 - \|z - z'\|_C^2 - 2(z' - z_0, z - z')_C \leq \|z - z_0\|_C^2. \end{aligned}$$

■

Proposition 6.3 *Suppose a_t, $t \in \mathbb{N}$ is a non-decreasing sequence of positive numbers such that*

$$\sum_{t=1}^{\infty} \frac{1}{a_t} < \infty.$$

Then

$$\sum_{t=1}^{\infty} \left[\frac{a_{t+1} - a_t - 1}{a_t} \right]^+ = +\infty.$$

Proof. Since

$$\sum_{t=1}^{\infty} \left[\frac{a_{t+1} - a_t - 1}{a_t} \right]^+ \geq \sum_{t=1}^{\infty} \frac{a_{t+1} - a_t}{a_t} - \sum_{t=1}^{\infty} \frac{1}{a_t}$$

and the last series converges, it is sufficient to show that

$$\sum_{t=1}^{\infty} \frac{a_{t+1} - a_t}{a_t} = +\infty.$$

Note that for $b \geq a > 0$, we have

$$\frac{b - a}{a} = \int_a^b \frac{1}{a}\, d\tau \geq \int_a^b \frac{1}{\tau}\, d\tau = \ln b - \ln a.$$

So,

$$\sum_{t=1}^{N} \frac{a_{t+1} - a_t}{a_t} \geq \sum_{t=1}^{N} (\ln a_{t+1} - \ln a_t) = \ln a_{N+1} - \ln a_1 \to +\infty \quad \text{as } N \to \infty. \quad ■$$

Lemma 6.4 *Suppose $\{\alpha_t\}$ is a sequence of real $m \times 1$ column vector, $I_t = \mathbf{I} + \sum_{s=1}^{t} \alpha_s \alpha_s^T$ diverges and κ_t is a sequence of positive numbers satisfying:*

$$I_t / \kappa_t \to G,$$

where G is a finite positive definite $m \times m$ matrix. Then

$$\sum_{t=N}^{\infty} \frac{1}{\kappa_t^{\delta}} \alpha_t^T I_t^{-1} \alpha_t < \infty$$

for any $\delta > 0$.

Proof. Since $tr(I_t) = m + \sum_{s=1}^{t} \alpha_s^T \alpha_s$ is a non-decreasing sequence of positive numbers, we have (see Proposition A2 in Sharia (2007))

$$\sum_{t=1}^{\infty} \frac{\alpha_t^T \alpha_t}{[tr(I_t)]^{1+\delta}} < \sum_{t=1}^{\infty} \frac{\alpha_t^T \alpha_t}{(\sum_{s=1}^{t} \alpha_s^T \alpha_s)^{1+\delta}} < \infty.$$

Since I_t/κ_t converges, we have that $tr(I_t)/\kappa_t$ tends to a finite limit, and

$$\sum_{t=1}^{\infty} \frac{\alpha_t^T \alpha_t}{\kappa_t{}^{1+\delta}} = \sum_{t=1}^{\infty} \frac{\alpha_t^T \alpha_t}{tr(I_t)^{1+\delta}} \left[\frac{tr(I_t)}{\kappa_t} \right]^{1+\delta} < \infty$$

Finally, since G_t is positive definite and we have $\kappa_t I_t^{-1} \to G^{-1}$, and it follows that $\kappa_t \lambda_t^{max}$ converges to a finite limit, where λ_t^{max} is the largest eigenvalue of I_t^{-1}. Thus,

$$\sum_{t=1}^{\infty} \frac{1}{\kappa_t^{\delta}} \alpha_t^T I_t^{-1} \alpha_t \leq \sum_{t=1}^{\infty} \frac{\alpha_t^T \alpha_t}{\kappa_t^{1+\delta}} \cdot \kappa_t \lambda_t^{max} < \infty. \quad \blacksquare$$

References

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

[2] Benveniste, A., Métivier, M., and Priouret, P. *Stochastic approximations and adaptive algorithms.* Springer-Verlag, 1990.

[3] Borkar, V. S. Stochastic approximation. *Cambridge Books* (2008).

[4] Campbell, K. Recursive computation of m-estimates for the parameters of a finite autoregressive process. *The Annals of Statistics* (1982), 442–453.

[5] Chen, H. F., Guo, L., and Gao, A.-J. Convergence and robustness of the robbins-monro algorithm truncated at randomly varying bounds. *Stochastic Processes and their Applications 27* (1987), 217–231.

[6] Chen, H. F., and Zhu, Y. M. Stochastic approximation procedures with randomly varying truncations. *Scientia Sinica Series A Mathematical Physical Astronomical & Technical Sciences 29*, 9 (1986), 914–926.

[7] Englund, J.-E., Holst, U., and Ruppert, D. Recursive estimators for stationary, strong mixing processesa representation theorem and asymptotic distributions. *Stochastic Processes and their Applications 31*, 2 (1989), 203–222.

[8] Fabian, V. On asymptotically efficient recursive estimation. *The Annals of Statistics* (1978), 854–866.

[9] Horn, R. A., and Johnson, C. R. Matrix analysis, 1985. *Cambridge, Cambridge.*

[10] Kallenberg, O. *Foundations of modern probability.* springer, 2002.

[11] Khasminskii, R. Z., and Nevelson, M. B. *Stochastic approximation and recursive estimation.* Nauka, Moscow, 1972.

[12] Kushner, H. J. Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 1 (2010), 87–96.

[13] Kushner, H. J., and Yin, G. *Stochastic approximation and recursive algorithms and applications*, vol. 35. Springer Science & Business Media, 2003.

[14] Lai, T. L. Stochastic approximation. *Annals of Statistics* (2003), 391–406.

[15] Lazrieva, N., Sharia, T., and Toronjadze, T. The robbins-monro type stochastic differential equations. i. convergence of solutions. *Stochastics: An International Journal of Probability and Stochastic Processes 61*, 1-2 (1997), 67–87.

[16] Lazrieva, N., Sharia, T., and Toronjadze, T. Semimartingale stochastic approximation procedure and recursive estimation. *Journal of Mathematical Sciences 153*, 3 (2008), 211–261.

[17] Lelong, J. Almost sure convergence of randomly truncated stochastic algorithms under verifiable conditions. *Statistics & Probability Letters 78*, 16 (2008), 2632–2636.

[18] Ljung, L., and Soderstrom, T. Theory and practice of recursive identification, 1987.

[19] Poljak, B. T., and Tsypkin, J. Z. Robust identification. *Automatica 16*, 1 (1980), 53–63.

[20] Robbins, H., and Monro, S. A stochastic approximation method. *The annals of mathematical statistics* (1951), 400–407.

[21] Robbins, H., and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In *Herbert Robbins Selected Papers*. Springer, 1985, pp. 111–135.

[22] Sakrison, D. J. Efficient recursive estimation; application to estimating the parameters of a covariance function. *International Journal of Engineering Science 3*, 4 (1965), 461–483.

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

[24] Sharia, T. On the recursive parameter estimation in the general discrete time statistical model. *Stochastic processes and their applications 73*, 2 (1998), 151–172.

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

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

[27] Sharia, T. Efficient on-line estimation of autoregressive parameters. *Mathematical Methods of Statistics 19*, 2 (2010), 163–186.

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

[29] Sharia, T. Truncated stochastic approximation with moving bounds: convergence. *Statistical Inference for Stochastic Processes* (2014), 1–17.

[30] Tadić, V. Stochastic gradient algorithm with random truncations. *European journal of operational research 101*, 2 (1997), 261–284.

[31] Tadić, V. Stochastic approximation with random truncations, state-dependent noise and discontinuous dynamics. *Stochastics: An International Journal of Probability and Stochastic Processes 64*, 3-4 (1998), 283–326.