Asymptotic Behaviour of Truncated Stochastic Approximation Procedures

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

We study asymptotic behaviour of stochastic approximation procedures with three main characteristics: truncations with random moving bounds, a matrix valued random step-size sequence, and a dynamically changing random regression function. In particular, we show that under quite mild conditions, stochastic approximation procedures are asymptotically linear in the statistical sense, that is, they can be represented as weighted sums of random variables. Therefore, a suitable form of the central limit theorem can be applied to derive asymptotic distribution of the corresponding processes. The theory is illustrated by various examples and special cases.

Keywords: Stochastic approximation, Recursive estimation, Parameter estimation

1 Introduction

This paper is the final part of the series of papers devoted to the study of truncated Stochastic approximation (SA) with moving bounds. The classical problem of SA is concerned with 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

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

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).)

In applications however, it is important to consider the setting when the function R changes over the time. So, let us assume that the objective now is to find a common root z^0 of a dynamically changing sequence of functions $R_t(z)$. Also, in certain circumstances it might be necessary to confine the values of the procedure to a certain set, or to a sequence of sets by applying a truncation operator. This happens if, e.g., the functions in the recursive equation are defined only for certain values of the parameter. 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 root z^0, e.g. a set, possibly time dependent, that contains the value of the unknown root. In order to study these procedures in an unified manner, we consider a SA of the following form

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

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, $U_t \subset \mathbb{R}^m$ is a random sequence of truncation sets, and Φ is the truncation operator which returns the procedure to U_t every time the updated value leaves the truncation set (see Section 2.1 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 (see Sharia (2014) for a more detailed discussions of these extensions).

Note that the idea of truncations 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)).

Convergence of the above class of procedures was studied in Sharia (2014) and the results on rate of convergence were established in Sharia and Zhong (2016). In this

paper, we derive further asymptotic properties of these procedures. In particular, we show that under quite mild conditions, SA procedures are asymptotically linear in the statistical sense, that is, they can be represented as weighted sums of random variables. Therefore, a suitable form of the central limit theorem can be applied to derive asymptotic distribution of the corresponding SA process. Since some of the conditions in the main statements might be difficult to interpret, we present explanatory remarks and corollaries. We also discuss the case of the classical SA and demonstrate that truncations with moving bounds make it possible to use SA even when the standard conditions on the function R do not hold. Finally, applications of the above results are discussed and some simulations are presented to illustrate the theoretical results of the paper. Proofs of some technical parts are postponed to Appendices.

2 Main results

2.1 Notation and preliminaries

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

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

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

Suppose that $h = h(z)$ is a real valued function of $z \in \mathbb{R}^m$. 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 $\mathbf{I}$. 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 **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 $\Phi_{U_t}(z)$ 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)).

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 $U = \emptyset$.*

2.2 Notes on convergence

Remark 2.2 This subsection contains simple results describing sufficient conditions for convergence and rate of convergence. We decided to present this material here 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) and Sharia and Zhong (2016).

Proposition 2.3 *Suppose that Z_t is a process defined by (2.1), U_t are admissible truncations for z^0.*
• *Suppose that*

(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.

- *Finally, if*

(W1)

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

eventually;

(W2) *there exist $0 < \delta \leq 1$ 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. See Remark 3.6 above.

2.3 Asymptotic linearity

In this subsection we establish that under certain conditions, the SA process defined by (2.1) is asymptotically linear in the statistical sense, that is, it can be represented as a weighted sum of random variables. Therefore, a suitable form of the central limit theorem can be applied to derive the corresponding asymptotic distribution.

Theorem 2.4 *Suppose that process Z_t is defined by* (2.1) *and*

(E1)

$$Z_t = Z_{t-1} + \gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})] \quad \textit{eventually.} \tag{2.4}$$

Suppose also that there exists a sequence of invertible random matrices A_t such that

(E2)

$$A_t^{-1} \longrightarrow 0 \quad \textit{and} \quad A_t \gamma_t(z^0) A_t \longrightarrow \eta \quad \textit{in probability,}$$

where $\eta < \infty$ (P-a.s.) is a finite matrix;

(E3)

$$\lim_{t\to\infty} A_t^{-1} \sum_{s=1}^{t} \left[\Delta\gamma_s^{-1}(z^0)\Delta_{s-1} + \tilde{R}_s(z^0 + \Delta_{s-1})\right] = 0$$

in probability, where

$$\Delta\gamma_s^{-1}(z^0) = \gamma_s^{-1}(z^0) - \gamma_{s-1}^{-1}(z^0),$$

$$\Delta_s = Z_s - z^0 \quad \textit{and} \quad \tilde{R}_s(z) = \gamma_s^{-1}(z^0)\gamma_s(z)R_s(z);$$

(E4)

$$\lim_{t\to\infty} A_t^{-1} \sum_{s=1}^{t} \left[\tilde{\varepsilon}_s(z^0 + \Delta_{s-1}) - \varepsilon_s(z^0)\right] = 0$$

in probability, where

$$\tilde{\varepsilon}_s(z) = \gamma_s^{-1}(z^0)\gamma_s(z)\varepsilon_s(z).$$

Then $A_t(Z_t - Z_t^*) \longrightarrow 0$ *in probability where*

$$Z_t^* = z^0 + \gamma_t(z^0) \sum_{s=1}^{t} \varepsilon_s(z^0);$$

that is, Z_t *is locally asymptotically linear in* z^0 *with* $\gamma_t = \gamma_t(z^0)$ *and* $\psi_t = \varepsilon_t(z^0)$.

Proof. Using the notation $\gamma_t = \gamma_t(z^0)$, $\varepsilon_t = \varepsilon_t(z^0)$ and $\Delta_t = Z_t - z^0$, (2.4) can be rewritten as

$$\Delta_t - \Delta_{t-1} = \gamma_t\tilde{R}_t(Z_{t-1}) + \gamma_t\tilde{\varepsilon}_t(Z_{t-1})$$

eventually. Multiplying both sides by γ_t^{-1}, we have

$$\sum_{s=1}^{t}[\gamma_s^{-1}\Delta_s - \gamma_{s-1}^{-1}\Delta_{s-1}] = \sum_{s=1}^{t}[\Delta\gamma_s^{-1}\Delta_{s-1} + \tilde{R}_s(Z_{s-1}) + \tilde{\varepsilon}_s(Z_{s-1})],$$

and since the sum on the left hand side reduces to $\gamma_t^{-1}\Delta_t - \gamma_0^{-1}\Delta_0$, we obtain

$$\Delta_t = \gamma_t \left[\mathcal{H}_t + \sum_{s=1}^{t} \tilde{\varepsilon}_s(Z_{s-1}) + \gamma_0^{-1}\Delta_0\right]$$

eventually, where

$$\mathcal{H}_t = \sum_{s=1}^{t}[\Delta\gamma_s^{-1}\Delta_{s-1} + \tilde{R}_s(Z_{s-1})].$$

Since $Z_t - Z_t^* = \Delta_t - (Z_t^* - z^0)$, we have

$$Z_t - Z_t^* = \gamma_t\Big[\mathcal{H}_t + \gamma_0^{-1}\Delta_0\Big] + \gamma_t \sum_{s=1}^{t}\Big[\tilde{\varepsilon}_s(Z_{t-1}) - \varepsilon_s\Big],$$

and

$$A_t(Z_t - Z_t^*) = A_t\gamma_t A_t A_t^{-1}\Big[\mathcal{H}_t + \gamma_0^{-1}\Delta_0\Big] + A_t\gamma_t A_t A_t^{-1}\sum_{s=1}^{t}\Big[\tilde{\varepsilon}_s(Z_{t-1}) - \varepsilon_s\Big]$$

eventually. By conditions (E2), (E3) and (E4), we have

$$A_t\gamma_t A_t \xrightarrow{P} \eta, \;\; A_t^{-1}\Big[\mathcal{H}_t + \gamma_0^{-1}\Delta_0\Big] \xrightarrow{P} 0 \quad \text{and} \quad A_t^{-1}\sum_{s=1}^{t}\Big[\tilde{\varepsilon}_s(Z_{t-1}) - \varepsilon_s\Big] \xrightarrow{P} 0$$

Therefore, $A_t(Z_t - Z_t^*) \longrightarrow 0$ in probability, that is, Z_t is locally asymptotically linear at z^0. ■

Proposition 2.5 *Suppose that A_t in Theorem 2.4 are positive definite diagonal matrices with non-decreasing elements and*

(Q1)

$$A_t^{-2}\sum_{s=1}^{t} A_s\Big[\Delta\gamma_s^{-1}(z^0)\Delta_{s-1} + \tilde{R}_s(z^0 + \Delta_{s-1})\Big] \longrightarrow 0$$

in probability, where $\tilde{R}_t$ is defined in (E3). Then (E3) in Theorem 2.4 holds.

Proof. Denote

$$\chi_s = A_s[\Delta\gamma_s^{-1}(z^0)\Delta_{s-1} + \tilde{R}_s(z^0 + \Delta_{s-1})]$$

and

$$A_t^{-1}\sum_{s=1}^{t}[\Delta\gamma_s^{-1}(z^0)\Delta_{s-1} + \tilde{R}_s(z^0 + \Delta_{s-1})] = A_t^{-1}\sum_{s=1}^{t} A_s^{-1}\chi_s\ .$$

Let us denote $P_s = A_s^{-1}$ and $Q_s = \sum_{m=1}^{s}\chi_m$. Then using the formula (summation by parts)

$$\sum_{s=1}^{t} P_s\Delta Q_s = P_t Q_t - \sum_{s=1}^{t}\Delta P_s Q_{s-1} \quad \text{with} \quad Q_0 = 0\ ,$$

we obtain

$$A_t^{-1}\sum_{s=1}^{t} A_s^{-1}\chi_s = A_t^{-2}\sum_{s=1}^{t}\chi_s + \mathcal{G}_t \quad \text{where} \quad \mathcal{G}_t = -A_t^{-1}\sum_{s=1}^{t}\Delta A_s^{-1}\sum_{m=1}^{s-1}\chi_m.$$

Since A_s are diagonal,

$$\Delta A_s^{-1} = A_s^{-1} - A_{s-1}^{-1} = -A_s^{-1}(A_s - A_{s-1})A_{s-1}^{-1} = -\Delta A_s A_s^{-1}A_{s-1}^{-1}.$$

Therefore,

$$\mathcal{G}_t = A_t^{-1}\sum_{s=1}^{t}\Delta A_s\left\{A_s^{-1}A_{s-1}^{-1}\sum_{m=1}^{s-1}\chi_m\right\}.$$

Denote by $A_s^{(j,j)}$ the j-th diagonal element of A_s. Since $0 \le A_{s-1}^{(j,j)} \le A_s^{(j,j)}$ for all j,

$$A_{s-1}^{-2}\sum_{m=1}^{s-1}\chi_m \longrightarrow 0 \implies A_s^{-1}A_{s-1}^{-1}\sum_{m=1}^{s-1}\chi_m \longrightarrow 0.$$

Because of the diagonality, we can apply the Toeplitz Lemma to the elements of $\mathcal{G}_t$, which gives

$$A_t^{-1}\sum_{s=1}^{t}[\Delta\gamma_s^{-1}(z^0)\Delta s - 1 + \tilde{R}_s(z^0+\Delta_{s-1})] = A_t^{-2}\sum_{s=1}^{t}\chi_s + \mathcal{G}_t \longrightarrow 0\,.$$

■

Proposition 2.6 *Suppose that A_t in Theorem 2.4 are positive definite diagonal matrices with non-decreasing elements. Denote by $\alpha^{(j)}$ the j-th element of $\alpha \in \mathbb{R}^m$ and by $A^{(j,j)}$ the j-th diagonal element of matrix A. Suppose also that*

(Q2)

$$\lim_{t\to\infty}(A_t^{(j,j)})^{-2}\sum_{s=1}^{t} E\Big\{\Big[\tilde{\varepsilon}_s^{(j)}(z^0+\Delta_{s-1}) - \varepsilon_s^{(j)}(z^0)\Big]^2\Big|\mathcal{F}_{s-1}\Big\} = 0$$

in probability P for all $j = 1,...,m$, where $\tilde{\varepsilon}_s$ is defined in (E4). Then (E4) in Theorem 2.4 holds.

Proof. Denote $M_t = \sum_{s=1}^{t}\Big[\tilde{\varepsilon}_s(z^0+\Delta_{s-1}) - \varepsilon_s(z^0)\Big]$. By the assumptions, M_t is a martingale and the quadratic characteristic $\langle M^{(j)}\rangle_t$ of the jth component $M_t^{(j)}$ is

$$\langle M^{(j)}\rangle_t = \sum_{s=1}^{t} E_{z^0}\Big\{\Big[\tilde{\varepsilon}_s^{(j)}(z^0+\Delta_{s-1}) - \varepsilon_s^{(j)}(z^0)\Big]^2\Big|\mathcal{F}_{s-1}\Big\}.$$

Using the Lenglart-Rebolledo inequality (see e.g., Liptser and Shiryayev (1989), Section 1.9), we have

$$P\Big\{(M_t^{(j)})^2 \geq K^2(A_t^{(j,j)})^2\Big\} \leq \frac{\epsilon}{K} + P\Big\{\langle M^{(j)}\rangle_t \ \ \geq \epsilon(A_t^{(j,j)})^2\Big\}$$

for each $K > 0$ and $\epsilon > 0$. Now by (Q2), $\langle M^{(j)}\rangle_t/(A_t^{(j,j)})^2 \longrightarrow 0$ in probability P and therefore $M_t^{(j)}/A_t^{(j,j)} \longrightarrow 0$ in probability P. Since A_t is diagonal, (E4) holds.■

Remark 2.7 Let us use Condition (E3) in Theorem 2.4 to construct an optimal step-size sequence $\gamma_t(z^0)$. Consider condition (Q1) in the one-dimensional case. Since $R_t(z^0) = 0$, we have

$$\begin{aligned} & A_t\left[\Delta\gamma_t^{-1}(z^0)\Delta_{t-1} + \tilde{R}_t(z^0+\Delta_{t-1})\right] \\ = & \left[\Delta\gamma_t^{-1}(z^0) + e_t\frac{R_t(z^0+\Delta_{t-1}) - R_t(z^0)}{\Delta_{t-1}}\right] A_t\Delta_{t-1}, \end{aligned}$$

where $e_t = \gamma_t^{-1}(z^0)\gamma_t(z^0+\Delta_{t-1})$. In most applications, the rate of A_t is $\sqrt{t}$ and $\sqrt{t}\Delta_t$ is stochastically bounded. Therefore, for (Q1) to hold, one should at least have the convergence

$$\Delta\gamma_t^{-1}(z^0) + e_t\frac{R_t(z^0+\Delta_{t-1}) - R_t(z^0)}{\Delta_{t-1}} \longrightarrow 0.$$

If $\gamma_t(z)$ is continuous, given that $\Delta_t \longrightarrow 0$, we expect $e_t \longrightarrow 1$. Therefore, we should have

$$\Delta\gamma_t^{-1}(z^0) \approx -R_t'(z^0).$$

Using the similar arguments for the multi-dimensional cases, we expect the above relation to hold for large t's, where $R_t'(z^0)$ is the matrix of the derivatives of $R_t(z)$ at $z = z^0$. So, an optimal choice of the step-size sequence should be

$$\gamma_t^{-1}(z) = -\sum_{s=1}^{t} R_s'(z),$$

or a sequence which is asymptotically equivalent to this sum.

Remark 2.8 (a) Condition (E1) in Theorem 2.4 holds if the truncations in (2.1) do not occur for large t's. More precisely, (E1) holds if for $t > T$ the truncations in (2.1) do not occur for some, possibly random T.
(b) Let us now consider the case when U_t is a shrinking sequence. For example, suppose that a consistent, but not necessarily efficient, auxiliary estimator $\tilde{Z}_t$ is

available. Then one can take the truncations on $U_t = S(\tilde{Z}_t, r_t)$, which is a sequence of closed spherical sets in $\mathbb{R}^m$ with the center at $\tilde{Z}_t$ and the radius $r_t \longrightarrow 0$. The resulting procedure is obviously consistent, as $\|Z_t - \tilde{Z}_t\| \leq r_t \longrightarrow 0$ and $\tilde{Z}_t \longrightarrow z^0$. However, if r_t decreases too rapidly, condition (E1) may fail to hold. Intuitively, it is quite obvious that we should not allow r_t to decreases too rapidly, as it may result in Z_t having the same asymptotic properties as $\tilde{Z}_t$, which might not be optimal. This truncation will be admissible if $\|\tilde{Z}_t - z^0\| < r_t$ eventually. In these circumstances, (E1) will hold if the procedure generates the sequence Z_t which converges to z^0 faster than r_t converges to 0.

(c) The considerations described in (b) lead to the following construction. Suppose that an auxiliary estimator $\tilde{Z}_t$ has a convergence rate d_t, in the sense that d_t is a sequence of positive r.v.'s such that $d_t \longrightarrow \infty$ and $d_t(\tilde{Z}_t - z^0) \to 0 \quad P$-a.s. Let us consider the following truncation sets

$$U_t = S\left(\tilde{Z}_t, c(d_t^{-1} + a_t^{-1})\right),$$

where c and a_t are positive and $a_t \longrightarrow \infty$. Then the truncation sequence is obviously admissible since $\|\tilde{Z}_t - z^0\| < cd_t^{-1}$ eventually. Now, if we can claim (using Proposition 2.3 or otherwise) that $a_t\|Z_t - z^0\| \longrightarrow 0$, then condition (E1) holds. Indeed, suppose that (E1) does not hold, that is, the truncations in (2.1) occur infinitely many times on a set A of positive probability. This would imply that Z_t appears on the surface of the spheres U_t infinitely many times on A. Since $z^0 \in S(\tilde{Z}_t, cd_t^{-1})$ eventually, we obtain that $\|Z_t - z^0\| \geq ca_t^{-1}$ infinitely many times on A, which contradicts our assumptions.

Another possible choice of the truncation sequence is

$$U_t = S\left(\tilde{Z}_t, c\left(d_t^{-1} \vee a_t^{-1}\right)\right).$$

(Here, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$). If we can claim by Proposition 2.3 or otherwise that $a_t\|Z_t - z^0\| \to 0$, then condition (E1) holds. Indeed, suppose that (E1) does not hold, that is, on a set A of positive probability the truncations in (2.1) occur infinitely many times. This would imply that

$$\|\tilde{Z}_t - Z_t\| = c(d_t^{-1} \vee a_t^{-1})$$

and

$$1 = c^{-1}(d_t \wedge a_t)\|\tilde{Z}_t - Z_t\| \leq c^{-1}(d_t \wedge a_t)\|\tilde{Z}_t - z^0\| + c^{-1}(d_t \wedge a_t)\|Z_t - z^0\|$$

infinitely many times on A, which contradicts our assumptions.

3 Special models and examples

3.1 Classical problem of stochastic approximation

Consider the classical problem of stochastic approximation to find a root z^0 of the equation $R(z^0) = 0$. Note that in the classical case, the step-size sequence can in general be of the form form $\gamma_t(Z_{t-1}) = a_t^{-1}\gamma(Z_{t-1})$. However, without loss of generality we can assume that $\gamma_t = a_t^{-1}\mathbf{I}$, since $\gamma(Z_{t-1})$ can be included in R and ε_t. Therefore, taking the step-size sequence $\gamma_t = a_t^{-1}\mathbf{I}$, where $a_t \longrightarrow \infty$ is a predictable scalar process, let us consider the procedure

$$Z_t = \Phi_{U_t}\Big(Z_{t-1} + a_t^{-1}[R(Z_{t-1}) + \varepsilon_t(Z_{t-1})]\Big). \tag{3.1}$$

Remark 3.1 In the corollary below we derive simple sufficient conditions for asymptotic linearity in the case when $a_t = t$. We also assume, using Proposition 2.3 or otherwise, that $t^{\delta/2}(Z_t - z^0) \longrightarrow 0$ for any $\delta \in (0,1)$. Note also that the condition (A1) below requires that the procedure is designed in such a way that the truncations in (3.1) do not occur for large t's (see Remark 2.8 for a detailed discussion of this requirement).

Corollary 3.2 *Suppose that Z_t is defined by (3.1), $a_t = t$ and $t^{\delta/2}(Z_t - z^0) \longrightarrow 0$ for any $\delta \in (0,1)$. Suppose also that*

(A1)

$$Z_t = Z_{t-1} + \frac{1}{t}[R(Z_{t-1}) + \varepsilon_t(Z_{t-1})] \quad \textit{eventually;}$$

(A2)

$$R(z^0 + u) = -u + \alpha(u) \quad \textit{where} \quad \|\alpha(u)\| = O(u^{1+\epsilon})$$

as $u \to 0$ for some $\epsilon > 0$;

(A3)

$$t^{-1}\sum_{s=1}^{t} E\Big\{\Big[\varepsilon_s(z^0 + u_s) - \varepsilon_s(z^0)\Big]^2 \Big| \mathcal{F}_{s-1}\Big\} < \infty,$$

where u_s is any predictable process with the property $u_s \longrightarrow 0$.

Then Z_t is asymptotically linear.

Proof. Let $A_t = \sqrt{t}\mathbf{I}$, then $A_t\gamma_t A_t = \mathbf{I}$ since $\gamma_t = \mathbf{I}/t$. Condition (E2) in Theorem 2.4 is satisfied. On the other hand, since $\tilde{R}(z) = R(z)$ and $\Delta\gamma_t^{-1} = \mathbf{I}$, we have

$$A_t^{-2}\sum_{s=1}^{t} A_s\left[\Delta\gamma_s^{-1}\Delta_{s-1} + \tilde{R}_s(Z_{s-1})\right] = \frac{1}{t}\sum_{s=1}^{t}\sqrt{s}[\Delta_{s-1}+R(z^0+\Delta_{s-1})] = \frac{1}{t}\sum_{s=1}^{t}\sqrt{s}\alpha(\Delta_{s-1}).$$

By (A2), there exists a constant $K > 0$ such that

$$\|\sqrt{s}\alpha(\Delta_{s-1})\| \le K\left\|\sqrt{s}\Delta_{s-1}^{1+\epsilon}\right\| = K\left\|\sqrt{\frac{s}{s-1}}\left[(s-1)^{\frac{1}{2(1+\epsilon)}}\Delta_{s-1}\right]^{1+\epsilon}\right\|$$

eventually. Since $1/[2(1+\epsilon)] < 1/2$, we have $(s-1)^{1/[2(1+\epsilon)]}\Delta_{s-1} \longrightarrow 0$, and therefore $\|\sqrt{s}\alpha(\Delta_{s-1})\| \longrightarrow 0$ as $\Delta_s \longrightarrow 0$. Thus, by the Toeplitz Lemma (see Lemma 5.1 in Appendix A),

$$\frac{1}{t}\sum_{s=1}^{t}\sqrt{s}\alpha(\Delta_{s-1}) \longrightarrow 0$$

So, (Q2) in Proposition 2.5 holds implying that condition (E3) in Theorem 2.4 is satisfied. Since $\tilde{\varepsilon}_t(z) = \varepsilon_t(z)$, it follows from (A3) that condition (Q2) in Proposition 2.6 holds. This implies that (E4) in Theorem 2.4 holds. Thus, all the conditions of Theorem 2.4 hold, implying that Z_t is asymptotically linear. ■

Remark 3.3 Using asymptotic linearity, the asymptotic normality is an immediate consequence of Corollary 3.2. Indeed, we have $\sqrt{t}(Z_t - Z_t^*) \longrightarrow 0$ in probability, where

$$Z_t^* = z^0 + \frac{1}{t}\sum_{s=1}^{t}\varepsilon_s(z^0).$$

So, Z_t and Z_t^* have the same asymptotic distribution. Now, to obtain the asymptotic distribution of Z_t, it remains only to apply the central limit theorem for martingales.

Remark 3.4 Note that condition (A2) above assumes that R function should be scaled in such a way that the derivative at z^0 is -1. Alternatively, a step-size sequence should be considered of the form $\gamma_t(Z_{t-1}) = t^{-1}\gamma(Z_{t-1})$, with appropriately chosen $\gamma(Z_{t-1})$. Detailed discussion of selection of an appropriate step-size sequence in the context of statistical parametric estimation is given in Section 3.3.

Example 3.5 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

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

Unless $l = 1$, we cannot use the standard SA without truncations as the standard condition on the rate of growth at infinity does not hold. So, we consider Z_t defined by (3.1) with a slowly expanding truncation sequence $U_t = [-u_t, u_t]$, where

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

We can assume for example, that $u_t = Ct^{r/2l}$, 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 for simplicity that the measurement errors are state free with the property that $\sum_{t=1}^{\infty} \sigma_t^2 a_t^{-2} < \infty$, where $\sigma_t^2 = E\{\varepsilon_t^2 \mid \mathcal{F}_{t-1}\}$. Then $|Z_t - z^0|$ converges (P-a.s.) to a finite limit. Furthermore, if z^0 is a unique root, then $Z_t \longrightarrow z^0$ (P-a.s.) provided that $\sum_{t=1}^{\infty} a_t^{-1} = \infty$. Finally, if Z_t is defined by (3.1) with $a_t = C_1 t$, then $t^{\alpha}(Z_t - z^0) \xrightarrow{a.s.} 0$ for any $\alpha < 1/2$ (see Sharia and Zhong (2016) for details). So, it follows that conditions in Corollary 3.2 hold (with R replaced by $C_1^{-1}R$), implying that Z_t is locally asymptotically linear. Now, depending on the nature of the error terms, one can apply a suitable form of the central limit theorem to obtain asymptotic normality of Z_t.

3.2 Linear procedures

Consider the recursive procedure

$$Z_t = Z_{t-1} + \gamma_t(h_t - \beta_t Z_{t-1}) \tag{3.2}$$

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 (3.2) 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

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

Remark 3.6 Recursive procedures (3.2) are linear in the sense that they locate the common root z^0 of the linear functions $R_t(u) = \beta_t(z^0 - u)$. The second part of the corollary below shows that the process Z_t is asymptotically linear in the statistical sense, that is, it can be represented as a weighted sum of random variables. The first part of the corollary below contains sufficient conditions for convergence and rate of convergence. We decided to present this material here for the sake of completeness, noting that the proof can be found in Sharia and Zhong (2016) (note also that (G1) below will hold if, e.g., $\Delta\gamma_t^{-1} = \beta_t$).

Corollary 3.7 *Suppose that Z_t is defined by (3.2) with $E(h_t|\mathcal{F}_{t-1}) = \beta_t z^0$ and a_t is a non-decreasing positive predictable process.*

1. *Suppose that*

(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.).

2. *Suppose that $\gamma_t \longrightarrow 0$ and*

$$\gamma_t^{1/2} \sum_{s=1}^{t} (\Delta\gamma_s^{-1} - \beta_s)\Delta_{s-1} \longrightarrow 0 \tag{3.3}$$

in probability, where $\Delta_t = Z_t - z^0$.

Then Z_t is asymptotically linear, that is,

$$\gamma_t^{1/2}(Z_t - z^0) = \gamma_t^{-1/2} \sum_{s=1}^{t} \varepsilon_s + r_t(z^0),$$

where $r_t(z^0) \longrightarrow 0$ in probability.

Proof. Let us check the conditions of Theorem 2.4 for $A_t = \gamma_t^{-1/2}$. Conditions (E1) and (E2) trivially hold. Since $\varepsilon_t(u) = h_t - \beta_t z^0$ is state free (i.e. does not depend on u), (E4) also holds. Since $\tilde{R}_t(Z_{t-1}) = R_t(Z_{t-1}) = -\beta_t\Delta_{t-1}$, we have

$$A_t^{-1} \sum_{s=1}^{t} \Big(\Delta\gamma_s^{-1}(z^0)\Delta_{s-1} + \tilde{R}_s(z^0 + \Delta_{s-1})\Big) = \gamma_t^{1/2} \sum_{s=1}^{t} (\Delta\gamma_s^{-1} - \beta_s)\Delta_{s-1}\ ,$$

and (E3) now follows from (3.3). Thus, all conditions of Theorem 2.4 are satisfied which implies the required result. ■

Example 3.8 Corollary 3.7 can be applied to study asymptotic behaviour of recursive least squares estimators in regression or time series models. To demonstrate this, let us consider a simple example of AR(1) process

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

where ξ_t is a sequence of square integrable random variables with mean zero. Consider the recursive least squares (LS) estimator of θ defined by

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

where $\hat{\theta}_0$ and $\hat{I}_0 > 0$ are any starting points and $\hat{I}_t = \hat{I}_0 + \sum_{s=1}^{t} X_{s-1}^2$. This procedure is clearly a particular case of (3.2) with

$$z^0 = \theta, \qquad Z_t = \hat{\theta}_t, \qquad \gamma_t = \hat{I}_t^{-1}, \qquad h_t = X_{t-1} X_t, \qquad \beta_t = X_{t-1}^2.$$

Since $\Delta \gamma_t^{-1} = X_{t-1}^2 = \beta_t$, condition (G1) holds (see Corollary 5.2 in Sharia and Zhong (2016)). Also, since

$$h_t - \beta_t z^0 = X_{t-1}(X_t - X_{t-1}\theta) = X_{t-1}\xi_t,$$

it follows that

$$E\{(h_t - \beta_t z^0)^T \gamma_t (h_t - \beta_t z^0) | \mathcal{F}_{t-1}\} = X_{t-1}^2 \hat{I}_t^{-1} E\{\xi_t^2 | \mathcal{F}_{t-1}\}.$$

Let $0 < \delta < 1$. Then taking $a_t = \hat{I}_t^{\delta}$ in (G2) we obtain

$$\sum_{t=1}^{\infty} a_t^{-1} E\{(h_t - \beta_t z^0)^T \gamma_t (h_t - \beta_t z^0) | \mathcal{F}_{t-1}\} = \sum_{t=1}^{\infty} \frac{1}{\hat{I}_t^{1+\delta}} X_{t-1}^2 E\{\xi_t^2 | \mathcal{F}_{t-1}\}$$

Now, since $\Delta \hat{I}_t = X_{t-1}^2$, if $\hat{I}_t \to \infty$ then the sum above is finite even if the conditional variances $E\{\xi_t^2 | \mathcal{F}_{t-1}\}$ go to infinity with rate $\hat{I}_t^{\delta^0}$, as far as $\delta^0 < \delta$ (this trivially follows from, e.g., Lemma 6.3 in Sharia and Zhong (2016)).

Let us now assume for simplicity that ξ_t is a sequence of i.i.d. r.v.'s with mean zero and variance 1. Then consistency and rate of convergence follows without any further moment assumptions on the innovation process. Indeed, since $\hat{I}_t \to \infty$ for

any $\theta \in \mathbb{R}$ (see, e.g, Shiryayev (1984, Ch.VII, §5), it follows that all the conditions of part 1 in Corollary 3.7 hold implying that $I_t^{1+\delta}(\hat{\theta}_t - \theta)^2$ converges a.s. to a finite limit for any $0 < \delta < 1$ and $\theta \in \mathbb{R}$.

Furthermore, since $\Delta\gamma_t^{-1} = \beta_t$, (3.3) trivially holds. It therefore follows that $\hat{\theta}_t$ is asymptotically linear and asymptotic normality is now obtained by applying the central limit theorem for i.i.d. random variables.

3.3 Application to parameter estimation

Let $X_1, \ldots, X_n$ be random variables with a joint distribution depending on an unknown parameter θ. Then an M-estimator of θ is defined as a solution of the estimating equation

$$\sum_{i=1}^{n} \psi_i(\theta) = 0, \tag{3.4}$$

where $\psi_i(\theta) = \psi_i(X_1^i; \theta)$, $i = 1, 2, \ldots, n$, are suitably chosen functions which may, in general, depend on the vector $X_1^i = (X_1, \ldots, X_i)$ of all past and present observations. If $f_i(x, \theta) = f_i(x, \theta | X_1, \ldots, X_{i-1})$ is the conditional probability density function or probability function of the observation X_i, given $X_1, \ldots, X_{i-1}$, then one can obtain a MLE (maximum likelihood estimator) on choosing

$$\psi_i(\theta) = l_t(\theta) = [f_i'(\theta, X_i | X_1^{i-1})]^T / f_i(\theta, X_i | X_1^{i-1}). \tag{3.5}$$

Besides MLEs, the class of M-estimators includes estimators with special properties such as robustness. Under certain regularity and ergodicity conditions, it can be proved that there exists a consistent sequence of solutions of (3.4) which has the property of local asymptotic linearity.

Let us consider estimation procedures which are recursive in the sense that each successive estimator is obtained from the previous one by a simple adjustment. In particular, we consider a class of estimators

$$\hat{\theta}_t = \Phi_{U_t}\Big[\hat{\theta}_{t-1} + \gamma_t(\hat{\theta}_{t-1})\psi_t(\hat{\theta}_{t-1})\Big], \qquad t \geq 1,$$

where ψ_t is a suitably chosen vector process, γ_t is a matrix valued step-size process, and $\hat{\theta}_0 \in \mathbb{R}^m$ is an initial value. This type of recursive estimators are especially convenient when the corresponding ψ-functions are non-linear in θ and therefore, solving (3.4) would require a numerical method (see e.g., Example 3.9). A detailed discussion and a heuristic justification of this estimation procedure are given in Sharia (2008).

The above procedure can be rewritten in the SA form. Indeed, assume that θ is an arbitrary but fixed value of the parameter and denote

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

Following the argument in Remark 2.7 (see also Sharia (2010)), an optimal step-size sequence would be

$$\gamma_t^{-1}(\theta) = -\sum_{s=1}^{t} R_s'(\theta)$$

If $\psi_t(z)$ is differentiable w.r.t. z and differentiation of $R_t(z) = E_\theta\{\psi_t(z) \mid \mathcal{F}_{t-1}\}$ is allowed under the integral sign, then $R_t'(z) = E_\theta\{\psi_t'(z) \mid \mathcal{F}_{t-1}\}$. This implies that, for a given sequence of estimating functions $\psi_t(\theta)$, another possible choice of the step-size sequence is

$$\gamma_t(\theta)^{-1} = -\sum_{s=1}^{t} E_\theta\{\psi_s'(\theta) \mid \mathcal{F}_{s-1}\},$$

or any sequence with the increments

$$\triangle\gamma_t^{-1}(\theta) = \gamma_t^{-1}(\theta) - \gamma_{t-1}^{-1}(\theta) = -E_\theta\{\psi_t'(\theta) \mid \mathcal{F}_{t-1}\}.$$

Also, since $\psi_t(\theta)$ is typically a P^θ-martingale difference,

$$0 = \int \psi_t(\theta, x \mid X_1^{t-1}) f_t(\theta, x \mid X_1^{t-1}) \mu(dx),$$

and if the differentiation w.r.t. θ is allowed under the integral sign, then (see Sharia (2010) for details)

$$E_\theta\{\psi_t'(\theta) \mid \mathcal{F}_{t-1}\} = -E_\theta\{\psi_t(\theta) l_t^T(\theta) \mid \mathcal{F}_{t-1}\},$$

where $l_t(\theta)$ is defined in (3.5). Therefore, another possible choice of the step-soze sequence is any sequence with the increments

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

Therefore, since the process

$$M_t^\theta = \sum_{s=1}^{t} \psi_s(\theta)$$

is a P^θ-martingale, the above sequence can be rewritten as

$$\gamma_t^{-1}(\theta) = \langle M^\theta, U^\theta \rangle_t$$

where $U_t^\theta = \sum_{s=1}^t l_s(\theta)$ is the score martingale.

Let us consider a likelihood case, that is $\psi_t(\theta) = l_t(\theta)$, the above sequence is the conditional Fisher information

$$I_t(\theta) = \sum_{s=1}^{t} E\{l_s(\theta) l_s^T(\theta) | \mathcal{F}_{s-1}\}. \tag{3.6}$$

Therefore, the corresponding recursive procedure is

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} + I_t^{-1}(\hat{\theta}_{t-1}) l_t(\hat{\theta}_{t-1})\Big), \qquad t \geq 1, \tag{3.7}$$

Also, given that the model possesses certain ergodicity properties, asymptotic linearity of (3.7) implies asymptotic efficiency. In particular, in the case of i.i.d. observations, it follows that the above recursive procedure is asymptotically normal with parameters $(0,\ i^{-1}(\theta))$, where $i(\theta)$ is the one-step Fisher information.

3.3.1 The i.i.d case

Consider the classical scheme of i.i.d. observations $X_1, X_2, ...$ having a common probability density function $f(x, \theta)$ w.r.t. some σ- finite measure μ, where $\theta \in \mathbb{R}^m$. Suppose that $\psi(x, \theta)$ is an estimating function with

$$E_\theta \{\psi(X_1, \theta)\} = \int \psi(x, \theta) f(x, \theta) \mu(dx) = 0.$$

A recursive estimator $\hat{\theta}_t$ can be defined by

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} + a_t^{-1} \gamma(\hat{\theta}_{t-1}) \psi(X_t, \hat{\theta}_{t-1})\Big)$$

where a_t is a non-decreasing real sequence, $\gamma(\theta)$ is an invertible $m \times m$ matrix and truncation sequence U_t is admissible for θ. In most applications $a_t = t$ and an optimal choice of $\gamma(\theta)$ is

$$\gamma(\theta) = \Big[E_\theta\Big\{\psi(X_t, \theta) l^T(X_t, \theta)\Big\}\Big]^{-1} \quad \text{where} \quad l(x, \theta) = \frac{[f'(x, \theta)]^T}{f(x, \theta)}\ .$$

Example 3.9 *Let $X_1, X_2, \ldots$ be i.i.d. random variables from 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. Denote*

$$\log'\mathbf{\Gamma}(\theta) = \frac{d}{d\theta}\log\mathbf{\Gamma}(\theta), \quad \log''\mathbf{\Gamma}(\theta) = \frac{d^2}{d\theta^2}\log\mathbf{\Gamma}(\theta).$$

Then

$$\frac{f'(x,\theta)}{f(x,\theta)} = \log x - \log'\mathbf{\Gamma}(\theta) \qquad \textit{and} \qquad i(\theta) = \log''\mathbf{\Gamma}(\theta),$$

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

$$\hat{\theta}_t = \Phi_{U_t}\left(\hat{\theta}_{t-1} + \frac{1}{t\ \log''\mathbf{\Gamma}(\hat{\theta}_{t-1})}\left[\log X_t - \log'\mathbf{\Gamma}(\hat{\theta}_{t-1})\right]\right) \tag{3.8}$$

with $U_t = [\alpha_t, \beta_t]$ *where* $\alpha_t \downarrow 0$ *and* $\beta_t \uparrow \infty$ *are sequences of positive numbers. Then it can be shown that (see Appendix B) if*

$$\sum_{t=1}^{\infty}\frac{\alpha_{t-1}^2}{t} = \infty \qquad \textit{and} \qquad \sum_{t=1}^{\infty}\frac{\log^2\alpha_{t-1} + \log^2\beta_{t-1}}{t^2} < \infty, \tag{3.9}$$

then $\hat{\theta}_t$ *is strongly consistent and asymptotically efficient, i.e.,* $\hat{\theta}_t \xrightarrow{a.s.} \theta$ *as* $t \longrightarrow \infty$, *and*

$$\mathcal{L}\Big(t^{1/2}(\hat{\theta}_t - \theta)|P^{\theta}\Big) \xrightarrow{w} \mathcal{N}\Big(0, \log''\mathbf{\Gamma}(\theta)\Big).$$

For instance,

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

with some positive constants C_1 *and* C_2, *obviously satisfy* (3.9).

The above result can be derived by rewriting (3.8) *in the form of the stochastic approximation (see Appendix B for details), i.e.,*

$$\hat{\theta}_t = \Phi_{U_t}\left(\hat{\theta}_{t-1} + \frac{1}{t}\left[R(\hat{\theta}_{t-1}) + \varepsilon_t(\hat{\theta}_{t-1})\right]\right) \tag{3.10}$$

where

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

and

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

4 Simulations

4.1 Finding roots of polynomials

Let us consider a problem described in Section 3.5 with

$$R(z) = -(z - z^0)^7 + 2(z - z^0)^6 - 5(z - z^0)^5 - 3(z - z^0),$$

and suppose that the random errors are independent Student random variables with degrees of freedom 7. Consider SA procedure (3.1) with $a_t = 3t$ and the truncation sequence $U_t = [-\log 3t, \log 3t]$. Then (see Example 3.5), it follow that this procedure is consistent, i.e., converges almost surely to z^0, and asymptotically linear. Also, since the error terms are i.i.d., it follows that the procedure is asymptotically normal. Note that the SA without truncations fails to satisfy the standard condition on the rate of growth at infinity. Here, slowly expanding truncations are used to artificially slow down the growth of R at infinity.

Figure 1 shows 30 steps of the procedure with starting points at -2, 0 and 5 respectively, where the root $z^0 = 2$. A histogram of the estimator over 500 replications (with $Z_0 = 0$) is shown in Figure 2.

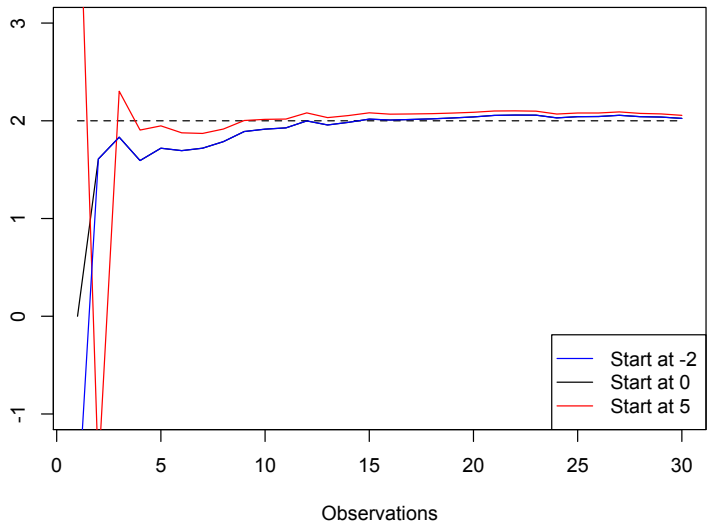

Figure 1: Realizations of the estimator in the polynomial example

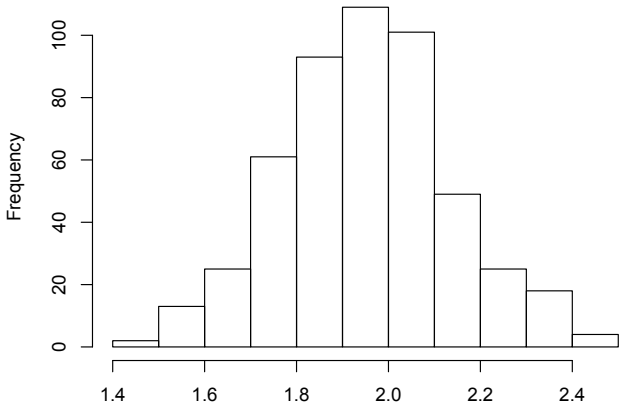

Figure 2: Histogram of the estimator in the polynomial example

4.2 Estimation of the shape parameter of the Gamma distribution

Let us consider procedure (3.8) in Example 3.9 with following two sets of truncations $U_t = [\alpha_t, \beta_t]$.

(1) FT – Fixed truncations: $\alpha_t = \alpha$ and $\beta_t = \beta$ where $0 < \alpha < \beta < \infty$.

(2) MT – Moving truncations: $\alpha_t = C_1[\log(t+2)]^{(-1/2)}$ and $\beta_t = C_2(t+2)$ where C_1 and C_2 are positive constants.

Figure 3 shows realizations of procedures (3.8) when $\theta = 0.1$ and the starting point $\hat{\theta}_0 = 1$, $C_1 = 0.1$, $C_2 = 1$ in MT, and $\alpha = 0.003$, $\beta = 100$ in FT. As we can see, the MT estimator approaches the true value of θ following a zigzag path. However, the FT estimator moves very slowly towards the true value of θ, caused by singularity at 0 of the functions appearing in the procedure.

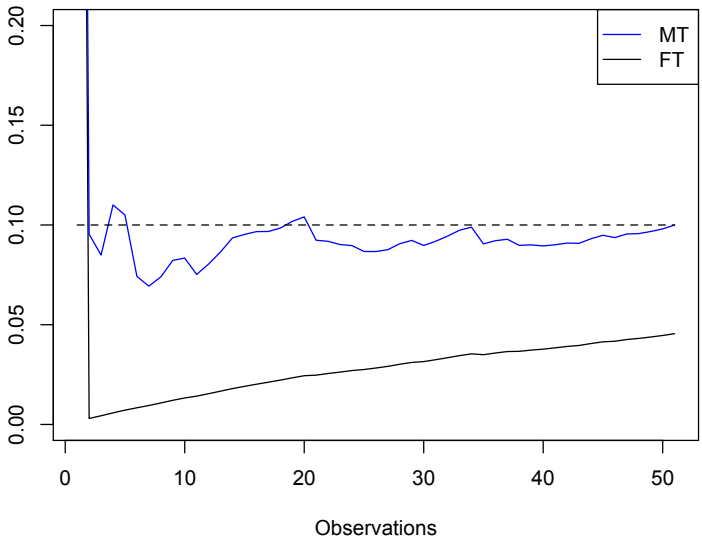

Figure 3: Performance of the estimator of the parameter in the Gamma distribution

5 Appendix

Lemma 5.1 *(The Toeplitz Lemma) Let* $\{a_n\}$ *be a sequence of non-negative numbers such that* $\sum_{n=1}^{\infty} a_n$ *diverges. If* $\nu_n \longrightarrow \nu_\infty$ *as* $n \longrightarrow \infty$*, then*

$$\lim_{n\longrightarrow\infty} \frac{\sum_{i=1}^{n} a_i\nu_i}{\sum_{i=1}^{n} a_i} = \nu_\infty \ .$$

Proof. Proof can be found in Loève (1977, P.250). ■

Properties of Gamma distribution In Example 3.9, we will need the following properties of the Gamma function (see, e.g., Whittaker (1927), 12.16). $\log'\mathbf{\Gamma}$ is increasing, $\log''\mathbf{\Gamma}$ is decreasing and continuous,

$$\log''\mathbf{\Gamma}(x) \leq \frac{1+x}{x^2}$$

and

$$\log''\mathbf{\Gamma}(x) \geq \frac{1}{x}. \tag{5.1}$$

Also (see Cramer (1946), 12.5.4),

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

Then,

$$E_\theta\left\{\log X_1\right\} = \log'\mathbf{\Gamma}(\theta), \qquad E_\theta\left\{(\log X_1)^2\right\} = \log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta))^2, \tag{5.2}$$

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

Using (5.1) and (5.2) we obtain

$$E_\theta\left\{\|R(u) + \varepsilon_t(u)\|^2 \mid \mathcal{F}_{t-1}\right\} = \frac{\log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u))^2}{(\log''\mathbf{\Gamma}(u))^2}. \tag{5.3}$$

The convergence to θ of the estimator defined by (3.8) is shown in Sharia (2014). To establish the rate of convergence, let us show that the conditions of Corollary 4.5 in Sharia and Zhong (2016) hold. Since

$$\begin{aligned} R'(u) = \frac{dR(u)}{du} &= -\frac{\log'''\mathbf{\Gamma}(u)}{\log''\mathbf{\Gamma}(u)} - \frac{\log'''\mathbf{\Gamma}(u)}{[\log''\mathbf{\Gamma}(u)]^2}\left(\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u)\right) \\ &= -1 - \frac{\log'''\mathbf{\Gamma}(u)}{[\log''\mathbf{\Gamma}(u)]^2}\left(\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u)\right), \end{aligned}$$

we have $R'(\theta) = -1 \leq -1/2$ and condition (B1) of Corollary 4.5 in Sharia and Zhong (2016) holds. Since $E_\theta\left\{\varepsilon_t(u) \mid \mathcal{F}_{t-1}\right\} = 0$, we have

$$E_\theta\left\{[R(u) + \varepsilon(u)]^2 \mid \mathcal{F}_{t-1}\right\} = R^2(u) + E_\theta\left\{\varepsilon_t^2(u) \mid \mathcal{F}_{t-1}\right\}. \tag{5.4}$$

Using (5.3) and (5.4),

$$\begin{aligned} E_\theta\left\{\varepsilon_t^2(u) \mid \mathcal{F}_{t-1}\right\} &\leq E_\theta\left\{[R(u) + \varepsilon(u)]^2 \mid \mathcal{F}_{t-1}\right\} \\ &= \log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u))^2, \end{aligned}$$

which is obviously a continuous function of u. Thus, for any $v_t \longrightarrow 0$, we have $E_\theta\left\{\varepsilon_t^2(\theta + v_t) \mid \mathcal{F}_{t-1}\right\}$ converges to a finite limit and so condition (BB) in Corollary 4.7 in Sharia and Zhong (2016) holds. Therefore, all the conditions of this corollary are satisfied with $a_t = t$ implying that $t^\delta(\hat{\theta}_t - \theta)^2 \xrightarrow{a.s.} 0$ for any $\delta < 1$.

Furthermore, since the second derivative of $R(u)$ exists, $R'(\theta) = -1$, and $R(\theta) = 0$, by the Taylor expansion,

$$R(\theta + u) = -u + R''(\tilde{u})u^2$$

for small u's and for some $\tilde{u} > 0$. Therefore, condition (A2) in Corollary 3.2 holds. It is also easy to check that

$$E_\theta\Big\{\Big[\varepsilon_s(\theta + u_s) - \varepsilon_s(\theta)\Big]^2 \Big| \mathcal{F}_{s-1}\Big\} \longrightarrow 0$$

for any predictable process $u_s \longrightarrow 0$. Condition (A3) is immediate from the Toeplitz Lemma. Thus, estimator $\hat{\theta}_t$ defined by (3.10) is asymptotic linear. Now, using the CLT for i.i.d. r.v.'s, it follows that $\hat{\theta}_t$ is asymptotically efficient.

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] 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.

[5] 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.

[6] Cramer, H. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946.

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

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

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

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

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

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

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

[14] Liptser, R., and Shiryayev, A. N. Theory of martingales. *Mathematics and its Applications. Kluwer, Dordrecht* (1989), 835–873.

[15] Loève, M. Probability theory. *Graduate texts in mathematics 45* (1977), 12.

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

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

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

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

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

[21] Sharia, T., and Zhong, L. Rate of convergence of truncated stochastic approximation procedures. *Mathematical Methods of Statistics (to appear)*.

[22] Shiryayev, A. N. Probability. *SpringerVerlag* (1984).

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

[24] 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.

[25] Whittaker, E. T., and Watson, G. N. *A course of modern analysis*. Cambridge university press, 1927.