Truncated Stochastic Approximation with Moving Bounds

PhD Thesis

Lei Zhong

Supervised by Dr. Teo Sharia

Department of Mathematics

Royal Holloway, University of London

Abstract

This thesis is concerned with a wide class of truncated stochastic approximation (SA) procedures. These procedures have three main characteristics: truncations with random moving bounds, a matrix valued random step-size sequence, and a dynamically changing random regression function. Convergence, rate of convergence, and asymptotic linearity of the SA procedures are established in a very general setting. Main results are supplemented with corollaries to establish different sets of sufficient conditions, with the main emphases on the parametric statistical estimation. The theory is illustrated by examples and special cases. Properties of these procedures are illustrated and discussed using a simulation study.

Acknowledgement

I am taking this opportunity to express my deepest gratitude and appreciation to my PhD supervisor, Dr Teo Sharia, who in spite of being extraordinarily busy with her duties, always takes time out to hear, guides and keeps me on the correct path. She teaches me not only how to do maths, but also how to think and write as a mathematician. This thesis can be completed as this final version only because of her insistence in a high standard and not giving up in the hardest time of my study.

I express my deepest thanks to Dr Alexey Koloydenko, my advisor, for taking part in useful decisions, giving necessary advices and guidance, and arranging all facilities to make life easier. I choose this moment to acknowledge his contribution gratefully.

Without the great support of my family, it would not be possible for me to write this thesis. I would like to thanks my mother Jinhua He and my wife Liuxuan Pan for their hard working in taking care of me and my babies, which let me be able to focus on my PhD study.

Special thanks to Prof Eugene Shargorodsky for his help and advice in the final stage of writing.

Contents

Chapter 1

Introduction

1.1 Introductory remarks

In the thesis, we deal with a large class of truncated Stochastic Approximation (SA) procedures with moving random bounds. 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. The main three features of the class of SA procedures considered here are: dynamically changing random regression functions, matrix-valued random step-size sequences, and truncations with random moving bounds.

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

$$Z_t = Z_{t-1} + \gamma_t \left(R(Z_{t-1}) + \varepsilon_t \right), \qquad t = 1, 2, \ldots \tag{1.1.1}$$

where ε_t is a sequence of zero-mean random variables and γ_t is a deterministic

sequence of positive numbers.

Recursion (1.1.1) is a classical Robbins-Monro Stochastic approximation (SA) procedure. Under certain conditions Z_t converges to the root z^0 of function R (see Section 1.2 for details). One of the most important applications of the above procedures is to that of the statistical parameter estimation.

Suppose that $X_1, \ldots, X_t$ are i.i.d. random variables and $f(x, \theta)$ is the common probability density function, where $\theta \in \mathbb{R}^m$ is an unknown parameter. Define a recursive estimation procedure for θ by

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{t} i(\hat{\theta}_{t-1})^{-1} \frac{f'^T(X_t, \hat{\theta}_{t-1})}{f(X_t, \hat{\theta}_{t-1})}, \qquad t \geq 1, \tag{1.1.2}$$

where $\hat{\theta}_0 \in \mathbb{R}^m$ is some starting value, and $i(\theta)$ is the one-step Fisher information matrix (f' denotes the row-vector of partial derivatives of f w.r.t. the components of θ). This estimator was introduced in Sakrison [62] and studied by a number of authors (see Section 1.2 for references). 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 (1.1.2) by rewriting it in the form of stochastic approximation with $\gamma_t = 1/t$,

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

Indeed, one can easily check that given certain standard assumptions, $R(\theta) = 0$ and ε_t is a martingale difference w.r.t. the filtration $\mathcal{F}_t$ generated by the observations.

So, the on-line recursive parameter estimation can be considered in the framework of the classical SA theory. However, the requirement of the i.i.d. model is too

restrictive in many applications.

Suppose now that we have a stochastic processes $X_1, X_2, \ldots$ 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 $\theta \in \mathbb{R}^m$ an unknown parameter. Then one can define (see detials in Section 3.1) the recursive estimator of θ by

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \gamma_t(\hat{\theta}_{t-1})\psi_t(\hat{\theta}_{t-1}), \qquad t \geq 1, \tag{1.1.3}$$

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

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

yields a likelihood type estimation procedure.

It turns out (see Sharia [69]) that to obtain an estimator with asymptotically optimal properties, a state-dependent matrix-valued random step-size sequences are needed. For the above procedure, a step size sequence $\gamma_t(u)$ with the property

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

is an optimal choice if we want to obtain on-line estimators with certain good asymptotic properties. For example, to derive a recursive procedure which has the same asymptotic properties as the maximum likelihood estimator (e.g., consistency and

asymptotic efficiency), we need to take

$$\psi_t(v) = \frac{[f_t'(X_t, v)]^T}{f_t(X_t, v)} \quad \text{and} \quad \gamma_t(v) = I_t^{-1}(v),$$

where $I_t(v)$ is the conditional Fisher information matrix (see Section 1.3). The recursion (1.1.3) can be written in the form of SA. Indeed, denote

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

where θ is arbitrary, but fixed value of the parameter. Then, $R_t(\theta) = 0$ for each t, and $\varepsilon_t(z)$ is a martingale difference.

In order to study these procedures in an unified manner, Sharia [70] introduced a SA of the following form

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

where $Z_0 \in \mathbb{R}^m$ is some starting value, $R_t(z)$ is predictable with the property that $R_t(z^0) = 0$ for all t's, $\gamma_t(z)$ is a matrix-valued predictable step-size sequence, $U_t \subset \mathbb{R}^m$ is a random sequence of truncation sets (see Section 2.1 for details).

Now, let us assume that at each step we have additional information about the root z^0. Lets us, e.g., assume that at each step t, $z^0 \in [\alpha_t, \beta_t]$, where α_t and β_t are random variables such that $-\infty < \alpha_t \leq \beta_t < \infty$. Then, one can consider a procedure, which at each step t produces approximations from interval $[\alpha_t, \beta_t]$. For example, a truncated classical SA procedure in this case can be derived using the following recursion

$$Z_t = \left[\ Z_{t-1} + \gamma_t \left(R(Z_{t-1}) + \varepsilon_t\right)\right]_{\alpha_t}^{\beta_t}, \qquad t = 1, 2, \ldots \tag{1.1.4}$$

where $[v]_a^b$ is the truncation operator, that is,

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

Truncated procedures may be useful in a number of circumstances. For example, if the functions in the procedure are defined only for certain values of the parameter, then the recursion 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 $\tilde{\theta}_t$ can be easily available having a rate d_t. Then to obtain asymptotically efficient estimator, we can construct a procedure with shrinking bounds by truncating the recursive procedure in a neighbourhood of θ with $[\alpha_t, \beta_t] = [\tilde{\theta}_t - \delta_t, \tilde{\theta}_t + \delta_t]$, $\delta_t \to 0$.

Thus, the SA procedures studied in the thesis 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. These procedures were introduced in Sharia [70]. Note that this approach is completely new as far as (1) and (2) are concerned. However, SA with truncations have been studied by various authors. (See Section 1.2 for comparison of this approach to the ones existing in the literature).

The thesis is organised as follows. Section 1.2 reviews some of the results in SA theory which are relevant to the thesis. Section 1.3 briefly describes general theory of parametric statistical estimation for discrete time stochastic processes. The main results of the thesis are given in Chapter 2. In particular, this chapter contains new results on the rate of convergence and asymptotic linearity of SA procedures under quite mild conditions. Also, a convergence result in Section 2.1 generalises the corresponding result in Sharia [70] by considering time dependent random Lyapunov type functions (see Lemma 2.2.1). This generalisation turns out to be quite useful as it allows to derive convergence results of the recursive parameter estimators in AR(m) models. Chapters 3 and 4 contain applications to parametric statistical estimation.

Section 2.3 contains new sets of sufficient conditions to derive the rate of convergence of SA procedures (see e.g., Lemma 2.3.7, Corollary 2.3.10 and Corollary 2.3.11). Section 2.4 contains new results on asymptotic linearity of SA procedures. In particular, this section establishes 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 a corre-

sponding SA process.

Section 2.5 also contains some new results (see Remark 2.5.10 and Section 2.6). Some results in Chapter 3 are new (e.g., Lemmas 3.2.4, 3.2.6 and Corollary 3.2.7). Chapter 4 contains application to the on-line parameter estimation in the autoregressive processes. The results presented in Sections 4.2 and 4.3 are also new, they generalize the corresponding results in one-dimensional case in Sharia [68].

Finally, in Chapter 5, some simulation results are presented to illustrate the theoretical results of the thesis.

Each chapter contains a brief introduction and a summary to explain novelty of the results presented in a given chapter. Main lemmas and theorems are followed by various corollaries and remarks containing sufficient conditions for the convergence and explaining some of the assumptions.

1.2 Robbins-Monro procedures

In 1951, Herbert Robbins and Sutton Monro introduced the basic stochastic approximation algorithms. Due to a large number of applications and the theoretical interests, their work attracted attention in the statistics literature immediately. After a series of important developments and improvements in the following 40 years, SA was developed as an important area of optimization and system control. It also became a powerful tool in many different fields including stochastic system control, recursive algorithms analysis and on-line parameter estimation. For the past 20 years, the SA type methods found many new applications, because of the growing need of on-line methods in a number of emerging disciplines. These areas broadly cover the adaptive control, signal processing, artificial neural networks and learning

algorithms.

The Robbins-Monro (RM) SA was developed to find a unique zero of a real valued function $R(z) : \mathbb{R} \to \mathbb{R}$ when only noisy measurements of R are available. In other words, one can only observe

$$Y_t(z) = R(z) + \xi_t, \tag{1.2.1}$$

where ξ_t are i.i.d. zero-mean random variables.

If the random noise ξ_t can be neglected and $R(x)$ is continuously differentiable, then the problem reduces to a standard problem of numerical analysis. However when only the noisy values of $R(z)$ are available, standard deterministic numerical methods do not work. In order to find an approximate value of z^0, Herbert Robbins and Sutton Monro [58] proposed a recursive procedure generated by

$$Z_t = Z_{t-1} + \gamma_t Y_t(Z_{t-1}) \tag{1.2.2}$$

where $\gamma_t > 0$ is a non-increasing real sequence. They asserted that Z_t converges to z^0 in the mean square sense under the following conditions:

(1) $(z - z^0)R(z) < 0$ for each $z \in \mathbb{R}\backslash\{z^0\}$;

(2) $\sum_{t=1}^{\infty} \gamma_t = \infty$ and $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$;

(3) $Y_t(z)$ is uniformly bounded.

It is easy to see that, on average, at each step the procedure moves towards the root z^0. Indeed, since $E\{\xi_t \mid \mathcal{F}_{t-1}\} = 0$, we have

$$E\left\{Z_t - Z_{t-1} \mid \mathcal{F}_{t-1}\right\} = \gamma_t R(Z_{t-1})$$

($\mathcal{F}_t$ is the σ-field generated by the random variables $Z_1, \ldots, Z_t$). If we suppose now that at time $t-1$, $Z_{t-1} < z^0$. Then, by condition (1), we have $E\{Z_t - Z_{t-1} \mid \mathcal{F}_{t-1}\} > 0$. So, the next step Z_t will be in the direction of z^0. If at time $t-1$, $Z_{t-1} > z^0$, then by the same reason, $E\{Z_t - Z_{t-1} \mid \mathcal{F}_{t-1}\} < 0$. So, on average, at each step the procedure moves towards z^0. Now, the second part of condition (2) ensures that the magnitude of the jumps $Z_t - Z_{t-1}$ decreases, so that Z_t does note oscillate around z^0 without approaching it. On the other hand, the first part of condition (2) is needed to guarantee that the jumps do not decrease too rapidly so that Z_t has enough "time" to reach z^0.

Robbins and Monro introduced their procedure in their seminal work published in 1951 which became one of the most cited articles in the history of Statistics. Later in 1952, inspired by this work, Kiefer and Wolfowitz [35] developed a SA method to find the maximum (or minimum) of a function, which is known as the Kiefer-Wolfowitz (KW) procedure.

The convergence of Robbins-Monro (RM) and KW procedures were initially proved in the mean square sense. Blum [9] established the almost sure (a.s.) convergence of RM and KW procedures under weaker assumptions. Dvoretzky [20] introduced a wider class of SA procedures, which allowed the study of RM and KW procedures in an unified manner.

Gladyshev [25] proved the convergence of RM procedures using a different approach, based on conditional expectation of $\|Z_t - z^0\|^2$ with respect to $\mathcal{F}_{t-1}$. Based on Gladyshev's work, Robbins and Siegmund [59] then established a general convergence theorem with martingale-difference measurement errors ξ_t.

To study the asymptotic normality of SA procedures, Burkholder [12], Hodges and Lehmann [31] and Chung [17] considered recursions for $E\{(Z_t - z^0)^k\}$ with

$k \geq 2$. They showed that under certain conditions, $\sqrt{t}(Z_t - z^0)$ is asymptotically normal. Sacks [61] proposed a different approach to study asymptotic behaviour by replacing the function R by its linear approximation in some neighbourhood of the root, and achieved the same results under weaker conditions. Lai and Robbins [42] (see also Lai and Robbins [43] and Wei [75]) generalized the asymptotic normality by adjusting the step-size sequences. These procedures are known as adaptive stochastic approximation.

Ljung [51] introduced the ordinary differential equation (ODE) method as a new way to analyse SA procedures. Using the fact that the effects of random noises average out asymptotically, Ljiung showed that the asymptotic behaviour could be described by certain ODEs. The ODE approach has become very popular, especially in the study of stability and asymptotic behaviour of SA procedures (see Kushner [37], Kushner and Clark [36] and Kushner and Shwartz [39]). This approach was also extensively explored in the seminal papers by Benmaim et al [4], [6] and [7].

Apart from the above approaches, a number of different techniques have been developed. For example, Delyon [19] asserted a purely deterministic method which was based on the deterministic conditions on the random noises. Polyak and Juditsky [56] proposed an averaging method, which accelerates Robbins-Monro procedure by using slower step-size sequences and averaging the values of the procedure for the last several steps. Other recent developments in SA theory can be found in Borkar [10], Lazrieva et al [47] and Benveniste et al [8].

Truncated SA procedures were studied in Hasminskii and Nevelson [28], Fabian [23], Chen and Zhu [16], Chen et al [15], Andradóttir [2] and Sharia [64], [68] and [70]. For example, an idea of truncations with shrinking bounds goes back to [28] and [23]. Truncations with expanding bounds were considered in [2] and also, in

the context of recursive parametric estimation, in [64] (see also [68]). Truncations with adaptive truncation sets of the Robbins-Monro SA were introduced in [15], and further explored and extended in Chen and Zhu [16], Andrieu, Moulines and Priouret [3], Tadić [72] and [73] and Lelong [49]. The latter algorithms are designed in such a way, that the procedure is pulled back to a certain pre-specified point or a set, every time the sequence leaves the truncation region. Truncation procedures considered in [70] are different from the latter ones and are similar to the the ones introduced in [28], [23] and [2]. A detailed comparison of these two different approaches can be found in [2].

Applications of SA can be found in a very wide range of fields. It plays an important role in adaptive control, machine learning and queueing theory. For example, Haykin [29] and Benaim [5] used SA method into learning approximation in neural networks. Watkins and Dayan [74] developed Q-learning and applied SA type procedures in Markov decision problems.

Due to its recursive nature, SA methods are naturally applied to construct on-line parameter estimation procedures. In 1965, Sakrison [62] proposed recursive SA type versions of the MLEs. In the i.i.d. models, these estimators were developed by Nevelson and Has'minskii [54], Fabian [23] and Poljak and Tsypkin [55]. Asymptotic behaviour of this type of SA procedures for non-i.i.d. models was studied by Campbell [13], Englund, Holst and Ruppert [21], Ljung and Soderstrom [50], Lai and Ying [45] and Sharia [65], [66], [67] and [68].

1.3 General statistical model

Let X_t, $t = 1, 2, \ldots$, be observations taking values in a measurable space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ equipped with a σ-finite measure μ. Suppose that the distribution of the process X_t depends on an unknown parameter $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^m$. Suppose also that for each $t = 1, 2, \ldots$, there exist regular conditional probability densities (see e.g. in Sharia [70]) of X_t given past observations $X_1, X_2, \ldots, X_{t-1}$, which will be denoted by

$$f_t(\theta, x_t \mid x_1^{t-1}) = f_t(\theta, x_t \mid x_1, \ldots, x_{t-1}),$$

where $x_1^{t-1} = (x_1, ..., x_{t-1})$ and $f_1(\theta, x_1 \mid x_1^0) = f_1(\theta, x_1)$ is a density function of the observation X_1. Note that the conditional probability densities are regular, if they correspond to the regular conditional distributions. The regularity is needed in order to write conditional expectations as integrals w.r.t. the corresponding conditional probability density functions (see e.g., Shiryayev [71], Theorem 2.7.3 and Definitions 2.7.4). Denote by $\mathcal{F}_t$ $(t = 1, 2, \ldots)$ the σ-field generated by the random variables $X_1, \ldots, X_t$, i.e.

$$\mathcal{F}_t = \sigma(X_1, \ldots, X_t).$$

There is no loss of generality in assuming that the basic space is the canonical space

$$(\Omega, \mathcal{F}) := (\mathbb{X}^\infty, \mathcal{B}(\mathbb{X}^\infty)),$$

where $\mathbb{X}^\infty = \{\mathbf{x} : \mathbf{x} = (x_1, x_2, \ldots),\ x_i \in \mathbb{X}\}$ and $\mathcal{B}(\mathbb{X}^\infty)$ is the σ-field generated by the cylindrical sets. Using Tulcea's theorem on extending a measure and the existence of a random sequence (see, e.g., Shiryayev [71], Ch.II, §9, Theorem 2), we

can construct the family

$$\left\{P^{\theta},\ \theta \in \Theta\right\}$$

of corresponding distributions on $(\mathbb{X}^{\infty}, \mathcal{B}(\mathbb{X}^{\infty}))$ and identify $X = (X_t)_{t=1,2,\ldots}$ with the coordinate process on $(\mathbb{X}^{\infty}, \mathcal{B}(\mathbb{X}^{\infty}))$, that is, $X_t(\mathbf{x}) = x_t,\ t = 1, 2, \ldots$.

Assume that $f_t(\theta, x_t \mid x_1^{t-1})$ is differentiable w.r.t. θ and denote by $f_t'(\theta)$ the row-vector of partial derivatives of f_t with respect to the components of θ, that is,

$$f_t'(\theta, x_t \mid x_1^{t-1}) = \frac{\partial}{\partial \theta} f_t(\theta, x_t \mid x_1^{t-1}) = \left(\frac{\partial f_t(\theta, x_t \mid x_1^{t-1})}{\partial \theta^{(1)}}, ..., \frac{\partial f_t(\theta, x_t \mid x_1^{t-1})}{\partial \theta^{(m)}} \right)$$

and

$$l_t(\theta, x_t \mid x_1^{t-1}) = \frac{[f_t'(\theta, x_t \mid x_1^{t-1})]^T}{f_t(\theta, x_t \mid x_1^{t-1})}$$

(with the convention $0/0 = 0$).

The *one-step Fisher information matrix* for $t = 1, 2, \ldots$ is defined as

$$i_t(\theta \mid x_1^{t-1}) = \int l_t(\theta, z \mid x_1^{t-1}) [l_t(\theta, z \mid x_1^{t-1})]^T f_t(\theta, z \mid x_1^{t-1}) \mu(dz).$$

We shall use the notation

$$f_t(\theta) = f_t(\theta, X_t \mid X_1^{t-1}), \qquad l_t(\theta) = l_t(\theta, X_t \mid X_1^{t-1}),$$

$$i_t(\theta) = i_t(\theta \mid X_1^{t-1}).$$

By definition, $i_t(\theta)$ is a version of the conditional expectation w.r.t. $\mathcal{F}_{t-1}$, i.e.

$$i_t(\theta) = E_{\theta} \left\{ l_t(\theta) [l_t(\theta)]^T \mid \mathcal{F}_{t-1} \right\}.$$

Everywhere in the present work conditional expectations are meant to be calculated as integrals w.r.t. the conditional probability densities.

The *Fisher information* at time t is

$$I_t(\theta) = \sum_{s=1}^{t} i_s(\theta).$$

The matrix $I_t(\theta)$ is a form of conditional information which reduces to the standard Fisher information in the case where the $X_t,\ \ t = 1, 2, \ldots$, are independent random variables.

An *estimator* (or *statistic*) based on t $(t = 1, 2, \ldots)$ observations is any $\mathcal{F}_t$ measurable random variable

$$\hat{\theta}_t = \hat{\theta}_t(X_1, \ldots, X_t).$$

An estimator is said to be *strongly consistent* if

$$P^\theta\{\lim_{t\to\infty} \hat{\theta}_t = \theta\} = 1$$

for each $\theta \in \Theta$.

Let for each $t = 1, 2, \ldots$

$$\psi_t(\theta, x_t, x_{t-1}, \ldots, x_1) \ : \ \ \Theta \times \mathbb{X}^t \ \ \mapsto \ \ \mathbb{R}^m$$

be Borel functions. Denote

$$\psi_t(\theta) = \psi_t(\theta, X_t, X_{t-1}, \ldots, X_1).$$

The process $\psi_t(\theta)(t = 1, 2, \dots)$ is said to be an *influence process* if

$$E_\theta \{\psi_t(\theta) \mid \mathcal{F}_{t-1}\} = 0 \tag{1.3.1}$$

(we assume that the conditional expectation in (1.3.1) is well-defined).

For fixed t, the function $\psi_t(\theta)$ is called an *influence function* (or *influence curve*).

Note that condition (1.3.1) means that the sequence

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

is a P^θ-martingale.

Note also that if differentiation of $f_t(\theta)$ is allowed under the integral sign, i.e.

$$0 = \frac{\partial}{\partial\theta} \int f_t(\theta, z \mid x_1^{t-1})\mu(dz) = \int f_t'(\theta, z \mid x_1^{t-1})\mu(dz)$$

then $l_t(\theta)$ is an influence process.

Suppose that $\psi_t(\theta)$ $(t = 1, 2, \dots)$ is an influence process. An *M-estimator* of θ is a solution of the equation

$$\sum_{s=1}^{t} \psi_s(\theta) = 0. \tag{1.3.2}$$

The choice $\psi_t(\theta) = l_t(\theta)$, yields the maximum likelihood estimator (MLE). Besides MLE-s, this class of estimators includes estimators constructed with special properties as robustness. Under certain regularity and ergodicity conditions (see e.g., Serfling [63], Huber [33] and Lehman and Casella [48]), it can be proved that there exists a consistent sequence of solutions of (1.3.2) which has the property of local

asymptotic linearity in the following sense.

We say that a discrete time stochastic process ξ_t $(t = 1, 2, ...)$ is *predictable*, if ξ_t is $\mathcal{F}_{t-1}$ measurable for each $t = 1, 2, ...$

Definition 1.3.1 An estimator $\hat{\theta}_t$ is said to be *locally asymptotically linear* if for each $\theta \in \Theta$ there exist an influence process $\psi_t(\theta)$ and a predictable process $\gamma_t(\theta)$ where $\det(\gamma_t(\theta)) \neq 0$ and $\gamma_t(\theta) \longrightarrow 0$ such that

$$\hat{\theta}_t = \theta + \gamma_t(\theta) \sum_{s=1}^{t} \psi_s(\theta) + K_t^{\theta},$$

and

$$A_t K_t^{\theta} \xrightarrow{P^{\theta}} 0,$$

where $A_t = A_t(\theta)$ is a predictable matrix valued stochastic process such that

$$\det(A_t) \neq 0, \quad A_t^{-1} \xrightarrow{P^{\theta}} 0 \quad \textit{and} \quad A_t \gamma_t A_t \xrightarrow{P^{\theta}} \eta$$

for some random matrix η and $t = 1, 2, ...$ (The symbol "$\xrightarrow{P^{\theta}}$" denotes the convergence in probability P^{θ}).

Asymptotic behaviour of an asymptotic linear estimator can be studied using suitable forms of the Central Limit Theorem (CLT) for martingales. For example, the next theorem, which is a simple corollary of the martingale CLT when $\Theta \subset \mathbb{R}$, shows that under certain conditions asymptotic linearity implies asymptotic normality.

Let $\mathcal{L}(\xi|P)$ denote the distribution of a random variable ξ w.r.t probability P, $\mathcal{N}(0, \sigma^2)$ is a Gaussian distribution with parameters $(0, \sigma^2)$ and "$\xrightarrow{w}$" denotes the weak convergence.

Theorem 1.3.2 *(Shiryayev [71], Ch. VII, §8, Theorem 4) Suppose that $\theta \in \mathbb{R}$ and for the influence process $\psi_t(\theta)$ there exists a non-random sequence of positive numbers $\gamma_t(\theta)$ such that*

(1)

$$j_t^{\psi}(\theta) := E_\theta \left\{ \psi_t^2(\theta) \mid \mathcal{F}_{t-1} \right\} < \infty,$$

(2)

$$\gamma_t(\theta) \sum_{s=1}^{t} j_s^{\psi}(\theta) \to \eta^{\psi}(\theta)$$

for some non-random $\eta^{\psi}(\theta) \geq 0$,

(3) *The Lindeberg condition holds, that is,*

for each $\varepsilon > 0$

$$\gamma_t(\theta) \sum_{s=1}^{t} E_\theta \left\{ \psi_s^2(\theta) \mathcal{I} \left\{ |\psi_s(\theta)| \geq \varepsilon \right\} \mid \mathcal{F}_{s-1} \right\} \xrightarrow{P^\theta} 0,$$

where $\mathcal{I}$ is the indicator function.

Then

$$\mathcal{L} \left(\gamma_t^{1/2}(\theta) \sum_{s=1}^{t} \psi_s(\theta) \mid P^\theta \right) \xrightarrow{w} \mathcal{N}(0, \eta^{\psi}(\theta)).$$

In the case of i.i.d. observations, influence functions usually depend only on θ and the current observation X_t, i.e.,

$$\psi_t(\theta) = \psi(\theta, X_t).$$

In this case, if $\psi_t(\theta)$ has a finite second moment

$$j^{\psi}(\theta) = j_t^{\psi}(\theta) = E_\theta \left\{ \psi_t^2(\theta) \right\}$$

and if $\gamma_t(\theta) = t^{-1}\gamma(\theta)$ for some non-random $\gamma(\theta)$, then conditions (1), (2) and (3) are trivially satisfied, and therefore

$$\mathcal{L}\left(t^{-1/2}\sum_{s=1}^{t}\psi_s(\theta) \mid P^\theta\right) \xrightarrow{\omega} \mathcal{N}(0, \gamma^2(\theta)j^\psi(\theta)). \tag{1.3.3}$$

In particular, if estimator $\hat{\theta}_t$ is locally asymptotically linear with

$$\psi_t(\theta) = l_t(\theta) = l(\theta, X_t)$$

and

$$\gamma_t(\theta) = I_t^{-1}(\theta) = [ti(\theta)]^{-1},$$

and if the likelihood function $l_t(\theta)$ has a finite second moment then it follows from (1.3.3) that $\hat{\theta}_t$ is asymptotically normal with parameters $(0, i^{-1}(\theta))$, i.e.

$$\mathcal{L}\left(t^{1/2}(\hat{\theta}_t - \theta) \mid P^\theta\right) \xrightarrow{\omega} \mathcal{N}(0, i^{-1}(\theta)),$$

that is, for large t, distribution of $\hat{\theta}_t$ can be approximated by $\mathcal{N}(\theta, (ti(\theta)^{-1}))$. Note that $I_t^{-1}(\theta) = [ti(\theta)]^{-1}$ is the Cramer-Rao lower bound for the variance of any unbiased estimator of θ. Details of martingale CLT in multi-dimensional cases can be found, e.g., in Heyde [30] (Theorem 12.6).

In the general situation we will use the following definition.

An estimator is said to be *asymptotically efficient* if it is asymptotically linear with

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

that is,

$$\hat{\theta}_t = \theta + I_t^{-1}(\theta) \sum_{s=1}^{t} l_s(\theta) + K_t^\theta$$

where $I_t^{-1} \longrightarrow 0$ and $I_t^{1/2} K_t^\theta \xrightarrow{P^\theta} 0$.

Sometimes this kind of efficiency is called asymptotically first order efficiency. The motivation behind this general definition is the same as in the classical scheme of i.i.d. observations. Under certain regularity and ergodicity conditions, $\sum_{s=1}^{t} l_s(\theta)$ is a P^θ-martingale and a martingale CLT can be applied to deduce that for large t, distribution of $\hat{\theta}_t$ can be approximated by $\mathcal{N}(\theta, I_t^{-1}(\theta))$. Note that asymptotic efficient estimators are not unique, as the same asymptotic distribution can be shared by many different estimators. However, under relatively mild conditions, for any asymptotic efficient estimators $\hat{\theta}_t$,

$$I_t^{1/2}(\hat{\theta}_t - T_t) \xrightarrow{P^\theta} 0$$

where T_t is the MLE (see, e.g., Hall and Heyde [27], Theorem 6.2).

That is, asymptotic efficient estimators are asymptotically equivalent to the MLE in the sense of leading to the same asymptotic distribution. For a detailed discussion of this notion see e.g., Hall and Heyde [27].

Chapter 2

Robbins-Monro Type Stochastic Approximation

This chapter contains main results of this thesis. Three main properties of the RM type SA are established: convergence, rate of convergence, and asymptotic linearity. The basic notations and conventions are given in Section 2.1. In Section 2.2, convergence and rate of convergence of the RM type SA are studied using the Robbins-Siegmund Lemma [59] and a sequence of the Lyapunov functions. The convergence result generalises the previous results in Sharia [70]. The result on the rate of convergence in Section 2.2 and asymptotic linearity in Section 2.4, are new in the field of SA. However, similar ways of analysis are developed in Sharia [66] and [69] to establish the corresponding properties in the parameter estimation context. Section 2.3 contains a number of corollaries, which help to verify conditions for the convergence and rate of convergence for specific statistical models in the following chapters. Finally, Section 2.5 contains discussion of the results of this chapter when applied to the classical SA problems.

2.1 Basic notions

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

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

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

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

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

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

$$h'(z) = \left(\frac{\partial}{\partial z^{(1)}}h(z), \ldots, \frac{\partial}{\partial z^{(m)}}h(z)\right).$$

Also, we denote by $h''(z)$ the matrix of second partial derivatives.

For any $a_t \in \mathbb{R}^m$ $(t = 1, 2, ...)$, we denote that $\Delta a_t = a_t - a_{t-1}$. Denote by $[a]^+$ and $[a]^-$ the positive and negative parts of $a \in \mathbb{R}$, i.e. $[a]^+ = \max(a, 0)$ and $[a]^- = \min(a, 0)$. The $m \times m$ identity matrix is denoted by $\mathbf{I}$.

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

$$Z_t = \Phi_{U_t}\Big(Z_{t-1} + \gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})]\Big), \quad t = 1, 2, \dots \tag{2.1.1}$$

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

$$E\left\{R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1}) \mid \mathcal{F}_{t-1}\right\} = R_t(Z_{t-1}), \tag{2.1.2}$$

$$E\left\{\varepsilon_t^T(Z_{t-1})\varepsilon_t(Z_{t-1}) \mid \mathcal{F}_{t-1}\right\} = \left[E\left\{\varepsilon_t^T(z)\varepsilon_t(z) \mid \mathcal{F}_{t-1}\right\}\right]_{z=Z_{t-1}}, \tag{2.1.3}$$

and the conditional expectations (2.1.2) and (2.1.3) are assumed to be finite. Here $\Phi_{U_t}(u)$ is the truncation operator such that

$$\Phi_{U_t}(z) = \begin{cases} z & \text{if } z \in U_t \\ z^* & \text{if } z \notin U_t, \end{cases} \tag{2.1.4}$$

where $z^* \in U_t$ minimizes the distance to z.

Denote $\Delta_t = Z_t - z^0$ where Z_t is defined by (2.1.1).

A random sequence of sets $U_t = U_t(\omega)$ is *admissible* for z^0 if (see Sharia [69] and [70])

• $U_t(\omega)$ is a closed convex subset of $\mathbb{R}^m$, for each t and ω;

• the truncation $\Phi_{U_t}(z)$ is $\mathcal{F}_t$ measurable, for each t and $z \in \mathbb{R}^m$;

• there exists $t_0(\omega) < \infty$ such that $z^0 \in U_t(\omega)$ whenever $t > t_0(\omega)$, for almost all ω (i.e., $z^0 \in U_t$ eventually).

One can use truncations which are based on the prior knowledge about the unknown root ($U_t = \mathbb{R}^m$ if there is no prior knowldge). Truncations may provide a simple tool to achieve an efficient use of information available in the estimation

process. Let us assume that a consistent, but not necessarily efficient auxiliary estimator $\tilde{Z}_t$ is available. Then one can use $\tilde{Z}_t$ to truncate the recursive procedure in a neighbourhood of z^0 by taking a 'spherical' $U_t = S(\tilde{Z}_t, d_t)$ with $\tilde{Z}_t$ as the center and $d_t \longrightarrow 0$ as the radius. That is,

$$\Phi_{U_t}(u) = \begin{cases} u & \text{if } \|u - \tilde{Z}_t\| \leq d_t \\ \tilde{Z}_t + \frac{d_t}{\|u - \tilde{Z}_t\|}(u - \tilde{Z}_t) & \text{if } \|u - \tilde{Z}_t\| > d_t. \end{cases}$$

Obviously, such a procedure is consistent. However, since the main goal is to construct an efficient estimator, care should be taken to ensure that the truncations do not shrink to z^0 too rapidly, otherwise Z_t will have the same asymptotic properties as $\tilde{Z}_t$.

The need of truncations may naturally arise from various reasons. One obvious consideration is that the functions in the procedure may only be defined for certain values of the parameter. In this case one would want the procedure to produce points only from this set $\Big($see Example 3.2.1$\Big)$. Truncations may also be useful when the standard assumptions such as restrictions on the growth rate of the relevant functions are not satisfied.

Remark 2.1.1 Note that (2.1.2) in fact means that the sequence $\varepsilon_t(Z_{t-1})$ is a martingale-difference. conditions (2.1.2) and (2.1.3) 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 Theorem 5.4 in Kallenberg [34]).

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. (For example, for random variables ξ and η, the relation $\xi < \eta$ means that $P(\xi < \eta) = 1$.)

• A sequence of random variables $(\zeta_t)_{t\geq 1}$ has some property ***eventually*** if for every ω in a set Ω_0 of P probability 1, the realisation $\zeta_t(\omega)$ has this property for all t greater than some $t_0(\omega) < \infty$.

• Assume that all conditional expectations are calculated as integrals w.r.t. corresponding regular conditional probability measures.

2.2 Convergence Lemmas

Lemma 2.2.1 *Suppose that Z_t is a process defined by* (2.1.1), (2.1.2) *and* (2.1.3). *Let $V_t(u) : \mathbb{R}^m \longrightarrow \mathbb{R}$ be a sequence of real valued non-negative functions having continuous and bounded partial second derivatives. Denote $\Delta_t = Z_t - z^0$ and $\Delta V_t(u) = V_t(u) - V_{t-1}(u)$. Suppose also 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

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

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

Furthermore, if truncation sequence U_t is admissible for z^0, conditions (V1) and (V2) hold, and there exists a set $A \in \mathcal{F}$ with $P(A) > 0$ such that for each $\epsilon \in (0,1)$

(V3)

$$\sum_{t=1}^{\infty} \inf_{\substack{\epsilon \le V_t(u) \le 1/\epsilon \\ z^0+u \in U_{t-1}}} [\mathcal{K}_t(u)]^- = \infty \quad \text{on } A, \tag{2.2.1}$$

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

Proof. Rewrite (2.1.1) 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,

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

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

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

we have

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

by using (2.1.2) and (2.1.3). Then, applying the obvious decomposition $\mathcal{K}_t = [\mathcal{K}_t]^+ - [\mathcal{K}_t]^-$, the previous inequality can be rewritten as

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

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

According to Lemma A.1 in Appendix A (with $X_t = V_t(\Delta_t), \beta_{t-1} = \xi_{t-1} = B_t$ and $\zeta_t = [\mathcal{K}_t(\Delta_{t-1})]^-$), by (V2) $\sum_{t=1}^{\infty} B_t < \infty$, it would imply that $V_t(\Delta_t)$ 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 (V3), this implies 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 2.2.2 *The conditions of the above Lemma are difficult to interpret. Therefore, the rest of the section and Section 2.3 are devoted to formulate lemmas and corollaries (Lemmas 2.2.5 and 2.3.7, Corollaries 2.3.1, 2.3.2, 2.3.4, 2.3.5, 2.3.10*

and 2.3.11) containing sufficient conditions for the convergence and the rate of convergence, and remarks (Remarks 2.2.3, 2.2.4, 2.3.6, 2.3.8, 2.3.9 and 2.3.12) explaining some of the assumptions. These corollaries are presented in such a way, that each subsequent corollary imposes conditions that are more restrictive than the previous one. For example, Corollary 2.3.11 and Remark 2.3.12 contains conditions which are most restrictive than all the previous ones, but are written in the simplest possible way.

Remark 2.2.3 Consider truncation sets $U_t = S(\alpha_t, r_t)$, where S denotes a closed sphere in $\mathbb{R}^m$ with center at $\alpha_t \in \mathbb{R}^m$ and radius r_t. Let $z_t' = \Phi_{U_t}(z_t)$ and suppose that $z^0 \in U_t$. Let C_t be 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 A.7 in Appendix A for details.) In particular, if C_t is a scalar matrix (i.e., $C_t = c\mathbf{I}$), condition (V1) automatically holds.

Remark 2.2.4 When condition (V1) holds, a typical choice of $V_t(u)$ is $V_t(u) = u^T C_t u$, where $\{C_t\}$ is a set of predictable positive semi-definite matrix process. Particularly, one can take C_t such that C_t/a_t goes to a finite matrix where $a_t \longrightarrow \infty$. Then, if conditions of Lemma 2.2.1 hold, we have $a_t \|Z_t - z^0\|^2$ tends to a finite limit and $Z_t \longrightarrow z^0$.

Lemma 2.2.5 *Suppose that (V1) and (V2) in Lemma 2.2.1 hold with $V_t(u) = V(u)$. Suppose also that truncation sequence U_t is admissible for z^0 and*

(L1) *for each $M > 0$,*

$$\inf_{\|u\| \geq M} V(u) > \delta > 0$$

for some δ;

(L2) *there exists a set* $A \in \mathcal{F}$ *with* $P(A > 0)$ *such that for each* $\epsilon \in (0, 1)$,

$$\sum_{t=1}^{\infty} \inf_{\substack{\epsilon \le V(u) \le 1/\epsilon \\ z^0 + u \in \bar{U}_{t-1}}} [\mathcal{N}_t(u)]^- = \infty \quad on \quad A,$$

where

$$\begin{aligned} \mathcal{N}_t(u) &= V'(u)\gamma_t(z^o + u)R_t(z^o + u) \\ &\quad + \frac{1}{2} \sup_v \|V''(v)\| E\Big\{ \|\gamma_t(z^o + u)\big[R_t(z^o + u) + \varepsilon_t(z^o + u)\big]\|^2 \mid \mathcal{F}_{t-1} \Big\}. \end{aligned} \tag{2.2.2}$$

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

Proof. Take $V_t(u) = V(u)$ in Lemma 2.2.1. condition (V3) follows from (L2) immediately, which implies that $V(\Delta_t) \longrightarrow 0$ (a.s.). Now, $\Delta_t \longrightarrow 0$ follows (L1) 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 $\|\Delta_{t_k}\| \ge \epsilon$ for some $\epsilon > 0$, and (2.2.5) would imply that $V(\Delta_{t_k}) > \delta > 0$ for large k-s, which contradicts the P-a.s. convergence $V(\Delta_t) \longrightarrow 0$. ■

2.3 Sufficient conditions for convergence and rate of convergence

Corollary 2.3.1 *Let* Z_t *be a process defined by* (2.1.1), (2.1.2) *and* (2.1.3). *Suppose that* U_t *are admissible truncations for* z^0, a_t *is a non-negative predictable scalar process and*

(C1) *for all* $z \in U_{t-1}$

$$\frac{\left[2(z-z^o)^T R_t(z) + a_t^{-1} E\left\{\|R_t(z)+\varepsilon_t(z)\|^2 \mid \mathcal{F}_{t-1}\right\}\right]^+}{1+\|z-z^o\|^2} \leq q_t \qquad (2.3.1)$$

eventually, where

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

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

Proof. Consider Lemma 2.2.1 with $V_t(u) = u^T u = \|u\|^2$ and the step-size sequence $\gamma_t(z) = a_t^{-1}\mathbf{I}$. Since U_t are admissible, condition (V1) holds. Also, we have $\Delta V_t(u) = 0$, $V'(u) = 2u^T$ and $V''(u) = 2\mathbf{I}$. Therefore,

$$\mathcal{K}_t(u) = 2u^T a_t^{-1} R_t(z^o+u) + a_t^{-2} E\left\{\|R_t(z^o+u)+\varepsilon_t(z^o+u)\|^2 \mid \mathcal{F}_{t-1}\right\}. \qquad (2.3.2)$$

Since $z^o + \Delta_{t-1} = Z_{t-1} \in U_{t-1}$,

$$\begin{aligned} & \frac{[\mathcal{K}_t(\Delta_{t-1})]^+}{1+V(\Delta_{t-1})} \\ = \; & a_t^{-1} \, \frac{\left[2\Delta_{t-1}^T R_t(z^o+\Delta_{t-1}) + \gamma_t E\left\{\|R_t(z^o+\Delta_{t-1})+\varepsilon_t(z^o+\Delta_{t-1})\|^2 \mid \mathcal{F}_{t-1}\right\}\right]^+}{1+\|\Delta_{t-1}\|^2} \\ \leq \; & a_t^{-1} q_t. \end{aligned}$$

Since $\sum_{t=1}^{\infty} q_t a_t^{-1} < \infty$ (P-a.s.), it follows from (C1) that condition (V2) hold. Therefore, $\|Z_t - z^0\|$ converges to a finite limit (P-a.s.). ■

Corollary 2.3.2 *Suppose that the conditions of Corollary 2.3.1 hold and*

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

$$\sum_{t=1}^{\infty} \inf_{\substack{\epsilon \le \|u\| \le 1/\epsilon \\ z^0+u \in U_{t-1}}} [\mathcal{N}_t(u)]^- = \infty, \qquad P\text{-a.s.},$$

where

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

Then $Z_t \longrightarrow z^o$ $(P\text{-a.s.})$, *for any initial value* Z_0.

Proof. Let us show that the conditions of Lemma 2.2.5 are satisfied with $V(u) = u^T u = \|u\|^2$ and $\gamma_t(z) = a_t^{-1}\mathbf{I}$. It follows from the proof of Corollary 2.3.1 that all the conditions of Lemma 2.2.1 hold with $V(u) = u^T u$. Hence, $\|Z_t - z^0\|$ converges. Since

$$\inf_{\|u\| \ge \epsilon} \|u\|^2 \ge \epsilon^2,$$

condition (L1) also trivially holds. Finally, (L2) is a consequence of (C2). Therefore, $Z_t \longrightarrow z^o$ $(P\text{-a.s.})$. ■

Remark 2.3.3 *Similar results of Corollary 2.3.1 and 2.3.2 can be found in Sharia [70].*

Corollary 2.3.4 *Suppose that* Z_t *is a process defined by* (2.1.1), (2.1.2) *and* (2.1.3), U_t *are admissible truncations for* z^0 *and*

(D1) *for large t's*

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

(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

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

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

Proof. Using condition (D1),

$$\begin{aligned} & \left[2(z - z^o)^T R_t(z) + a_t^{-1} E\left\{\|R_t(z) + \varepsilon(z)\|^2 \mid \mathcal{F}_{t-1}\right\}\right]^+ \\ \leq \;& a_t^{-1} E\left\{\|R_t(z) + \varepsilon(z)\|^2 \mid \mathcal{F}_{t-1}\right\} \end{aligned}$$

eventually. Hence conditions of Corollary 2.3.1 hold with $q_t = r_t a_t^{-1}$ and the result follows. ■

Corollary 2.3.5 *Suppose that the conditions of Corollary 2.3.4 are satisfied and*

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

$$\inf_{\substack{\epsilon \leq \|z - z^o\| \leq 1/\epsilon \\ z \in U_{t-1}}} -(z - z^0)^T R_t(z) > \nu_t \tag{2.3.3}$$

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. It follows from the poof of Corollary 2.3.4 that conditions of Corollary 2.3.1 hold. Let us prove that (C2) of Corollary 2.3.2 holds. Using the obvious inequality

$[a]^- \geq -a$, we have

$$[\mathcal{N}_t(u)]^- \geq -2u^T a_t^{-1} R(z^o + u) - a_t^{-2} E\left\{\|R_t(z^o + u) + \varepsilon_t(z^o + u)\|^2 \mid \mathcal{F}_{t-1}\right\}.$$

By conditions (D2) of Corollary 2.3.4 and taking the supremum of the conditional expectation above over the set $\{u : \epsilon \leq \|u\| \leq 1/\epsilon\}$, we obtain

$$\sup \frac{E\left\{\|R_t(z^o + u) + \varepsilon_t(z^o + u)\|^2 \mid \mathcal{F}_{t-1}\right\}}{1 + \|u\|^2} (1 + \|u\|^2) \leq r_t(1 + \|1/\epsilon\|^2).$$

Then, by (D3), taking the infimum over the same set,

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

Condition (C2) is now immediate from (D3) and (D2) of Corollary 2.3.4. Hence, according to Corollary 2.3.2, Z_t converges (a.s.) to z^0. ■

Remark 2.3.6 The rest of this section is concerned with the rate of convergence of (2.1.1). In most applications, checking conditions of Lemma 2.3.7 and Corollary 2.3.10 below is very 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). In this case, 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 Chapter 3).

Lemma 2.3.7 *Suppose that Z_t is a process defined by* (2.1.1), (2.1.2) *and* (2.1.3). *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 2.2.1 holds with $V_t(u) = u^T C_t u$ and

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

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

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 (a.s.).

Proof. Let us check the conditions of Lemma 2.2.1 with $V_t(u) = u^T C_t u$.

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

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

Because 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

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

and denoting

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

where

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

we have

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

By (R1), we have

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

Therefore,

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

where

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

Now, 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,

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

the condition (V2) of Lemma 2.2.1 has been satisfied. Thus, (V1) and (V2) hold, and $(Z_t - z^0)^T C_t(Z_t - z^0)$ converges to a finite limit (a.s.). ■

Remark 2.3.8 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 "high enough" absolute value, then it is possible to introduce a non-zero $\mathcal{P}_t$ without jeopardizing (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 2.3.9 The next corollary is a special case of Lemma 2.3.7 when the step-size matrix sequence, γ_t, is a sequence of scalar matrix, i.e. $\gamma_t(Z_{t-1}) = a_t^{-1}\mathbf{I}$, where a_t is a non-decreasing positive sequence.

Corollary 2.3.10 *Let Z_t be a process defined by* (2.1.1), (2.1.2) *and* (2.1.3) *with $\gamma_t(Z_{t-1}) = a_t^{-1}\mathbf{I}$, where $a_t > 0$ is a non-decreasing sequence. Suppose that truncation sequence U_t is admissible and*

(W1)

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

eventually;

(W2) *let $\delta < 1$,*

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

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

Proof. Consider Lemma 2.3.7 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

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

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 $(a_t/a_{t-1})^\delta \leq a_t/a_{t-1}$,

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

Therefore, (W1) leads to (R1). According to Remark 2.2.3, condition (V1) holds since $V_t(u) = a_t^\delta \|u\|^2$. Thus, all the conditions of Lemma 2.3.7 hold and $a_t^\delta \|Z_t - z^0\|^2$ converges to a finite limit (P-a.s.). ■

Corollary 2.3.11 *Let Z_t be a process defined by* (2.1.1), (2.1.2) *and* (2.1.3) *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) *$R_t'(z^0) \leq -1/2$ for large t's;*

(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 $|\sigma_t^2(\xi_t)| \leq K$ for large t's, 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 2.3.10 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 2.3.12 Corollary 2.3.11 gives simple (but more restrictive) sufficient conditions in one-dimensional cases to derive the rate of convergence. It is easy to see that all conditions of Corollary 2.3.11 trivially hold, if e.g., $R_t(z) = R(z)$ and ε_t are state independent i.i.d. random variables with a finite second moment.

2.4 Asymptotic linearity

In this section, we establish conditions for asymptotic linearity of process defined by (2.1.1). Once asymptotic linearity is established, one can use a suitable form of the CLT to derive asymptotic distribution of Z_t (see Section 1.3 for details).

Theorem 2.4.1 *Suppose that process Z_t is defined by* (2.1.1), (2.1.2), (2.1.3) *and*

(E1)

$$Z_t = Z_{t-1} + \gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})] \tag{2.4.1}$$

eventually.

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

(E2)

$$A_t^{-1} \longrightarrow 0$$

in probability and

$$A_t \gamma_t(z^0) A_t \longrightarrow \eta$$

in probability, where $\eta < \infty$ (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 $\Delta_t = Z_t - z^0$, (2.4.1) can be rewritten as

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

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

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

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

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

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

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

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. ■

Proposition 2.4.2 *Suppose that A_t in Theorem 2.4.1 are positive definite diagonal matrices with non-decreasing elements (i.e. $A_{t-1}^{(j,j)} \leq A_t^{(j,j)}$ for $j = 1...m$, where $A_t^{(j,j)}$ is the jth diagonal element of A_t) and*

(Q1)

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

in probability P, where $\tilde{R}_t$ is defined in (E3). Then (E3) in Theorem 2.4.1 holds.

Proof. Denote

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

then

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

Consider equation

$$\sum_{s=1}^{t} P_s \Delta Q_s = P_t Q_t - \sum_{s=1}^{t} \Delta P_s Q_{s-1} \quad \text{with} \quad P_0 Q_0 = 0$$

and let $P_s = A_s^{-1}$, $Q_s = \sum_{m=1}^{s} \chi_m$, then we obtain

$$A_t^{-1} \sum_{s=1}^{t} A_s^{-1}\chi_s = A_t^{-2} \sum_{s=1}^{t} \chi_s + \mathcal{G}_t \ ,$$

where

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

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,

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

Since $0 \leq A_{s-1}^{(j,j)} \leq A_s^{(j,j)}$ for all j (where $A_s^{(j,j)}$ is the j-th diagonal component of A_s),

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

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

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

Therefore, condition (E3) of Theorem 2.4.1 holds. ■

Proposition 2.4.3 *Suppose that A_t in Theorem 2.4.1 are positive definite diagonal matrices with non-decreasing elements. Denote by $\alpha^{(j)}$ the j-th component of $\alpha \in \mathbb{R}^m$ and by $A^{(j,j)}$ the j-th diagonal component of matrix A. Suppose also that*

(Q2)

$$E\Big\{ \tilde{\varepsilon}_s(z^0 + \Delta_{s-1}) - \varepsilon_s(z^0) \Big| \mathcal{F}_{s-1} \Big\} = 0;$$

(Q3)

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

in probability P for all $j=1,...,m$, where $\tilde{\varepsilon}_s$ is defined in (E4). Then (E4) in Theorem 2.4.1 holds.

Proof. Denote $M_t=\sum_{s=1}^{t}\Big[\tilde{\varepsilon}_s(z^0+\Delta_{s-1})-\varepsilon_s(z^0)\Big]$. By (Q2), M_t is a martingale. Then the quadratic characteristic $\langle M^{(j)}\rangle_t$ of martingale $M_t^{(j)}$ is

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

Use the Lenglart-Rebolledo inequality (see e.g., Liptser and Shiryayev [57], Section 1.9), we have

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

for each $K>0$ and $\epsilon>0$. Now by (Q3), $\langle M^{(j)}\rangle_t/(A_t^{(j,j)})^2\longrightarrow 0$ 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 2.4.4 Condition (E3) in Theorem 2.4.1 gives a useful suggestion for the optimal choice of the step-size sequence $\gamma_t(z^0)$. Consider condition (Q1) in the one-dimensional case. Since $R_t(z^0)=0$, we have

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

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 hold for large t's, where $R_t'(z^0)$ is the matrix of 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 asymptotically equivalent to this sum.

Remark 2.4.5 (a) Condition (E1) in Theorem 2.4.1 holds if the truncations in (2.1.1) do not occur for large t's. More precisely, (E1) holds if the truncations in (2.1.1) do not occur for $t > T$, 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 center at $\tilde{Z}_t$ and radius $r_t \longrightarrow 0$. Such a 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 if r_t decreases too rapidly, it may result in Z_t having the same asymptotic

properties as $\tilde{Z}_t$. 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 Lemma 2.3.7 or otherwise) that $a_t\|Z_t - z^0\| \longrightarrow 0$, then (E1) holds. Indeed, suppose that the truncations in (2.1.1) occur infinitely many times on a set A of positive probability. This would imply that Z_t appears on the surface of the spheres infinitely many times on A. Since $z^0 \in S(\tilde{Z}_t, cd_t^{-1})$, 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 Lemma 2.2.1 or otherwise that $a_t\|Z_t - z^0\| \to 0$, then (E1) holds. Indeed, suppose that on a set A of positive probability the truncations in (2.1.1) 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.

2.5 Classical problem of stochastic approximation

Consider the classical problem of stochastic approximation when function R is not dynamically changing with $R(z^0) = 0$. Let us take a step-size sequence $\gamma_t = a_t^{-1}\mathbf{I}$, where $a_t \longrightarrow \infty$ is a predictable scalar process ($a_t = t$ usually) and consider the procedure

$$Z_t = \Phi_{U_t}\Big(Z_{t-1} + a_t^{-1}[R(Z_{t-1}) + \varepsilon_t(Z_{t-1})]\Big). \tag{2.5.1}$$

Corollary 2.5.1 *Suppose that Z_t is a process defined by* (2.5.1)*, truncation sequence U_t is admissible, and*

(H1) for any $z \in \mathbb{R}^m$ with the property that $z \in U_t$ eventually,

$$(z - z^0)^T R(z) \leq 0;$$

(H2) there exists a predictable process p_t such that

$$\sup_{z \in U_{t-1}} \frac{\|R(z)\|^2}{1 + \|z - z^0\|^2} \leq p_t,$$

where

$$\sum_{t=1}^{\infty} a_t^{-2} p_t < \infty;$$

(H3) there exists a predictable process e_t such that

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

eventually where

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

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

Proof. Consider Corollary 2.3.4 with $R_t = R$. Condition (D1) holds immediately. Since $E\{\varepsilon_t(u) \mid \mathcal{F}_{t-1}\} = 0$, 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 $r_t = p_t + e_t$. Thus, by Corollary 2.3.4, $\|Z_t - z^0\|$ converges to a finite limit (a.s.). ■

Remark 2.5.2 Suppose that ε_t is an error term which does not depend on z and denote

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

Then condition (H3) holds if

$$\sum_{t=1}^{\infty} \sigma_t^2 a_t^{-2} < \infty, \qquad P\text{-a.s..} \tag{2.5.2}$$

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 2.5.3 *To compare the above result to that of Kushner-Clark 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 2.5.1 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 continuos 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 in Remark 2.5.2 will be sufficient for (H3).*

Remark 2.5.4 *A more general result of Corollary 2.5.1 can be found in Sharia [70] (Corollary 2.10).*

Corollary 2.5.5 *Suppose that the conditions of Corollary 2.5.1 are satisfied and*

(H4)

$$(z - z^0)^T R(z) < 0 \quad \textit{for all} \quad z \in U_t \backslash \{z^0\},$$

eventually;

(H5)

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

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

Proof. Consider Corollaries 2.3.4 and 2.3.5 with $\nu_t = \nu$. Conditions of Corollary 2.5.1 imply that conditions (D1) and (D2) hold (according to the proof of Corollary

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

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

eventually and by (H5)

$$\sum_{t=1}^{\infty} \nu a_t^{-1} = \nu \sum_{t=1}^{\infty} a_t^{-1} = \infty.$$

Thus, (D3) is satisfied. Therefore, $Z_t \longrightarrow z^0$ (a.s.). ■

Remark 2.5.6 If the truncation sets are bounded, then some of the conditions above can be weakened considerably. For example, condition (H2) in Corollary 2.5.1 will automatically hold given that $\sum_{t=1}^{\infty} a_t^{-2} < \infty$.

Also if it is only required that Z_t converges to a 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.

Corollary 2.5.7 *Suppose that Z_t, defined by* (2.5.1)*, converges to z^0 (a.s.) and truncation sequence U_t is admissible. Suppose also that*

(B1)

$$u^T R(z^0 + u) \le -\frac{1}{2}\|u\|^2$$

for small u's;

(B2)

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

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

$$\sum_{t=1}^{\infty} a_t^{\delta-2}\|R(z^0+v_t)\|^2 < \infty$$

and

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

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 (a.s.) to a finite limit.*

Proof. Since $\gamma_t = a_t^{-1}\mathbf{I}$ in this case, the result follows from Lemma 2.3.7 if we take $R_t = R$, $\rho_t = a_t^{-1}$, $\mathcal{P}_t = 0$ and $C_t = a_t^{\delta}\mathbf{I}$.

Here, $\lambda_t^{max} = \lambda_t^{min} = a_t^{\delta}$. Since $a_t \geq a_{t-1}$ eventually, $\delta < 1$ and by (B2),

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

for some constant C. So (B1) leads to (R1).

Since $Z_t \longrightarrow z^0$, by (B3) we have

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

(R2) has been met. Therefore, according to Lemma 2.3.7, $a_t^{\delta}\|Z_t - z^0\|^2$ converges (a.s.) to a finite limit. ■

Remark 2.5.8 It follows from Proposition A.11 in Appendix A 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 $\epsilon \leq 1$. Indeed,

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

Corollary 2.5.9 *Suppose that $Z_t \longrightarrow z^0$, where Z_t is defined by (2.5.1) with $a_t = t^{\epsilon}$ where $\epsilon \in (1/2, 1]$, truncation sequence U_t is admissible for z^0, and (B1) in Corollary 2.5.7 holds. Suppose also that R is continuous at z^0 and there exists $0 < \delta < 2 - 1/\epsilon$ such that*

(BB)

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

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 (a.s.).

Proof. Since $a_t = t^\epsilon$ where $\epsilon \in (1/2, 1]$, (B2) is satisfied. (See Remark 2.5.8)

Since R is continuous at z^0 and $Z_t \longrightarrow z^0$, $R(Z_{t-1})$ is bounded. Also, we have $(\delta-2)\epsilon < -1$, then

$$\sum_{t=1}^{\infty} a_t^{\delta-2}\|R(Z_{t-1})\|^2 = \sum_{t=1}^{\infty} t^{(\delta-2)\epsilon}\|R(Z_{t-1})\|^2 < \infty$$

and

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

as $v_t \longrightarrow 0$.

Thus, (B3) has been met. Now, condition (B1), (B2) and (B3) hold, according to Corollary 2.5.7, $t^\delta\|Z_t - z^0\|^2$ converges to a finite limit (a.s.). ■

Remark 2.5.10 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 2.5.9 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, whcih converge with the rate $t^{-1/3}$.

Corollary 2.5.11 *Suppose that Z_t is defined by* (2.5.1) *and $t^{\delta/2}(Z_t - z^0) \longrightarrow 0$ for any $\delta \in (0,1)$ and $a_t = t$. Suppose also that*

(A1)

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

eventually;

(A2)

$$R(z^0 + u) = -u + \alpha(u)$$

where

$$\|\alpha(u)\| = O(u^{1+\epsilon})$$

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 us check the conditions in Theorem 2.4.1. Condition (E1) follows from (A1). Let $A_t = \sqrt{t}\mathbf{I}$, then $A_t \gamma_t A_t = \mathbf{I}$ (note that $\gamma_t = 1/t$). Condition (E2) holds. On the other hand, since $\tilde{R}(z) = R(z)$,

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

There exists a constant $K > 0$ such that

$$\begin{aligned}\|\sqrt{s}\alpha(\Delta_{s-1})\| &\leq K \left\|\sqrt{s}(\Delta_{s-1})^{1+\epsilon}\right\| \\ &= K \left\|\sqrt{\frac{s}{s-1}}\left[(s-1)^{1/2(1+\epsilon)}\Delta_{s-1}\right]^{1+\epsilon}\right\|.\end{aligned}$$

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 Lamma A.4 in Appendix A),

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

and

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

According to Proposition 2.4.2, condition (E3) in Theorem 2.4.1 is satisfied. Since $\tilde{\varepsilon}_t(z) = \varepsilon_t(z)$ and (A3), conditions (Q2) and (Q3) in Proposition 2.4.3 hold. Therefore condition (E4) in Theorem 2.4.1 holds. Thus, all the conditions of Theorem 2.4.1 hold and Z_t is asymptotically linear. ■

Remark 2.5.12 Using asymptotic linearity, asymptotic normality is a immediate consequence of Corollary 2.5.11 in the classical SA. 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, it remains only to apply the CLT for martingales (see e.g. Theorem 1.3.2 for the one-dimensional

case).

Example 2.5.13 Let l be a positive integer and

$$R(z) = -\sum_{i=1}^{l} C_i (z - z^0)^i, \tag{2.5.3}$$

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

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

Consider a truncation sequence $U_t = [-u_t, u_t]$, where $u_t \longrightarrow \infty$ is a sequence of positive numbers. Suppose that

$$\sum_{t=1}^{\infty} a_t^{-1} = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} u_t^{2l}\ a_t^{-2} < \infty. \tag{2.5.4}$$

Then, provided that the measurement errors satisfy condition (H3) of Corollary 2.5.1, the truncated procedure (2.5.1) converges a.s. to z^0.

Indeed, condition (H1) of Corollary 2.5.1, (H4) and (H5) of Corollary 2.5.5 trivially hold. For large t's,

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

which implies condition (H2) of Corollary 2.5.1.

One can always choose a suitable sequence u_t which satisfies (2.5.4). For example, if the degree of the polynomial is known to be l (or at most l), and $a_t = 1/t$, 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 2.5.11 holds (e.g., one can assume for simplicity that $\varepsilon_t s$ 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 2.5.7 holds. The above convergence is a consequence of Corollary 2.5.9 and Remark 2.5.10.

Furthermore, conditions in Corollary 2.5.11 are satisfied if $a_t = C_1 t$ and the measurement errors are state free. Then Z_t is locally asymptotically linear. Now, depending on the nature of the error terms, one can apply a suitable form of the CLT to obtain the asymptotic normality of Z_t (e,g,. Theorem 1.3.2 if ε_t is martingale-difference).

Remark 2.5.14 A similar simpler example was considered by Chen [14]. However, our approach is different and is similar to that considered in Sharia [64] (see Section 1.2 for details). We plan to compare performances of these two approaches for the general polynomial functions in our future work.

2.6 Summary

In this chapter, some results are new in the field of SA. In particular, dynamically changing Lyapunov functions are used to verify the conditions for convergence. The convergence result given in this section generalises the corresponding result in Sharia [70] by considering time dependent random Lyapunov type functions (see Lemma 2.2.1). This generalisation turns out to be quite useful as it can be used to derive convergence results of the recursive parameter estimators in AR(m) models (see details in Chapter 4).

Note also that the proof of the convergence lemma (Lemma 2.2.1) is based on the Robbins-Siegmund Lemma (see Lemma A.1 in Appendix A). This lemma is, in fact, a special case of the theorem on the convergence sets of nonnegative semi-martingales (see, e.g., Lazrieva et al [46]). This observation might be useful if one wants to generalise the results of the thesis to the continuous time SA.

Sufficient conditions for the rate of convergence is derived in Lemma 2.3.7, Corollaries 2.3.10 and 2.3.11, which are also new.

The results in section 2.4, show that under quite mild conditions, the SA process 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. Furthermore, the results in this section help to identify step-size sequences that are optimal for a given set of R functions (see Remark 2.4.4). Corollaries 2.4.2 and 2.4.3 give sufficient conditions to derive asymptotic linearity by using the Toeplitz Lemma. The same idea is also used in Sharia [69]. These corollaries will be used in verifying asymptotic linearity of specific statistical models later in Chapters 3 and 4.

In Section 2.5, the above results are applied to the functions which are not dynamically changing against time t. In this case, the conditions are similar to those in the well known classical SA (see e.g., Lai and Robbins [43]). It shows that the conditions we used in the thesis are minimal in the sense that they do not impose any additional restrictions when applied to the classical case.

Furthermore, Section 2.5 contains new results even for the classical SA problem. In particular, truncations with moving bounds gives a possibility to use SA in the cases when classical conditions on the function R do not hold (see Example 2.5.13).

Also, Remark 2.5.10 highlights a very interesting link between the rate of the step-size sequence and the rate of convergence of the SA process in the classical case. This observation would not surprise experts working in this field. However, we failed to find it in a written form in the existing literature.

Chapter 3

Application to Parameter Estimation

This chapter is devoted to the application of the results established in the previous chapters to the problems of parametric statistical estimation. In i.i.d. models, the classical SA method can directly be applied to parametric statistical estimation problems and this approach has been exploited by a number of authors (see Section 1.2 for details). However, as it was mentioned in the introduction, to be able to apply SA to the parametric statistical problems in the general statistical model, one needs to consider generalisation presented in the previous chapters.

A class of recursive on-line SA type procedures for the general statistical model is introduced in Section 3.1. A brief comparison between the recursive estimation method and the Newton-Raphson type iterations is also given in this section.

Recursive estimation for specific models is discussed in Section 3.2, including the i.i.d. cases, the exponential family of Markov processes, and linear procedures.

3.1 Recursive on-Line estimation for the general statistical model

Let X_t, $t = 1, 2, \ldots$, be observations taking values in a measurable space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ equipped with a σ-finite measure μ. Suppose that the distribution of the process X_t depends on an unknown parameter $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^m$. Consider an estimating equation

$$\sum_{s=1}^{t} \psi_s(v) = 0, \tag{3.1.1}$$

where $\psi_s(v) = \psi_s(v, X_s, X_{s-1}, ..., X_1)$ is an influence process. Then an M-estimator $\hat{\theta}_t$ of θ is defined as a solution of equation (3.1.1) (see Section 1.3 for details). If $\psi_s(v)$ functions are linear w.r.t. v, then the estimator derived from (3.1.1) is naturally on-line. For example, if $X_1, X_2, \ldots, X_t$ are i.i.d. $N(\theta, \sigma^2)$ r.v.'s, then the likelihood equation is

$$\sum_{s=1}^{t} \frac{x_s - \theta}{\sigma^2} = 0$$

and the MLE is $\hat{\theta}_t = \bar{X}$, which can be written as

$$\frac{1}{t}\sum_{s=1}^{t} X_s = \frac{1}{t}\sum_{s=1}^{t-1} X_s + \frac{1}{t}X_t = \frac{t-1}{t}\frac{1}{t-1}\sum_{s=1}^{t-1} X_s + \frac{1}{t}X_t.$$

That is,

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{t}(X_t - \hat{\theta}_{t-1}). \tag{3.1.2}$$

So, the estimator $\hat{\theta}_t$ at each step t is obtained from the estimator at the previous step $\hat{\theta}_{t-1}$ and the new information X_t.

In general, to find a possible form of an approximate recursive relation, consider

$\hat{\theta}_t$ defined as a root of the estimating equation (3.1.1). Denoting the left hand side of (3.1.1) by $M_t(v)$ and assuming that the difference $\hat{\theta}_t - \hat{\theta}_{t-1}$ is "small" we can write $M_t(\hat{\theta}_t) \approx M_t(\hat{\theta}_{t-1}) + M'_t(\hat{\theta}_{t-1})(\hat{\theta}_t - \hat{\theta}_{t-1})$ and

$$0 = M_t(\hat{\theta}_t) - M_{t-1}(\hat{\theta}_{t-1}) \approx M'_t(\hat{\theta}_{t-1})(\hat{\theta}_t - \hat{\theta}_{t-1}) + \psi_t(\hat{\theta}_{t-1}).$$

Therefore,

$$\hat{\theta}_t \approx \hat{\theta}_{t-1} - \frac{\psi_t(\hat{\theta}_{t-1})}{M'_t(\hat{\theta}_{t-1})},$$

where $M'_t(\theta) = \sum_{s=1}^{t} \psi'_s(\theta)$. Now, depending on the nature of the underlying model, $M'_t(\theta)$ can be replaced by a simpler expression. For instance, in the i.i.d. models with $\psi(x, v) = f'(x, v)/f(x, v)$ (the MLE case), by the strong law of large numbers,

$$\frac{M'_t(\theta)}{t} = \frac{1}{t}\sum_{s=1}^{t} \left(f'(X_s, \theta)/f(X_s, \theta)\right)' \approx E_\theta \left[\left(f'(X_1, \theta)/f(X_1, \theta)\right)'\right] = -i(\theta)$$

for large t's, where $i(\theta)$ is the one-step Fisher information. So, in this case, one can consider

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{t\ i(\hat{\theta}_{t-1})} \frac{f'(X_t, \hat{\theta}_{t-1})}{f(X_t, \hat{\theta}_{t-1})}, \qquad t \geq 1, \tag{3.1.3}$$

to construct an estimator which is "asymptotically equivalent" to the MLE.

Note that the MLE in the i.i.d. normal case has exactly this form, indeed, (3.1.2) can trivially be rewritten as

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{1}{t} \underbrace{1/\sigma^2}_{i(\theta)} \underbrace{\frac{(X_t - \hat{\theta}_{t-1})}{\sigma^2}}_{\frac{f'(X_t, \hat{\theta}_{t-1})}{f(X_t, \hat{\theta}_{t-1})}}.$$

Motivated by the above argument, one can consider a class of estimators

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} + \gamma_t(\hat{\theta}_{t-1})\psi_t(\hat{\theta}_{t-1})\Big), \qquad t \geq 1, \tag{3.1.4}$$

where ψ_t is a suitably chosen vector process, γ_t is a step-size matrix process, and $\hat{\theta}_0 \in \mathbb{R}^m$ is some initial value. In particular, if $U_t = \mathbb{R}^m$ and $\psi_s(\theta) = f'_s(X_s, \theta)/f_s(X_s, \theta)$, where $f_s(x, \theta) = f_s(x, \theta | X_1, \ldots, X_{s-1})$ is the conditional pdf of the observation X_s given $X_1, \ldots, X_{s-1}$, we obtain

$$\hat{\theta}_t = \hat{\theta}_{t-1} + I_t^{-1}(\hat{\theta}_{t-1})\frac{f_t'^T(X_t, \hat{\theta}_{t-1})}{f_t(X_t, \hat{\theta}_{t-1})}, \qquad t \geq 1, \tag{3.1.5}$$

where, $I_t(\theta)$ is the conditional Fisher information matrix, f'_t is the row-vector of partial derivatives of f_t w.r.t. the components of θ.

It should be noted that recursions (3.1.3) and (3.1.5) resemble the Newton-Raphson or the one-step Newton-Raphson iterative procedures. In the i.i.d. case, the Newton-Raphson iteration for the likelihood equation is

$$\vartheta_k = \vartheta_{k-1} + J^{-1}(\vartheta_{k-1}) \sum_{s=1}^{t} \frac{f'(X_s, \vartheta_{k-1})}{f(X_s, \vartheta_{k-1})}, \qquad k \geq 1, \tag{3.1.6}$$

where $-J(v)$ is the second derivative of the log-likelihood function, that is,

$$\sum_{s=1}^{t} \frac{\partial}{\partial v} \left(f'(X_s, v)/f(X_s, v)\right)$$

or its expectation, that is, $-ti(v)$. In the latter case, the iterative scheme is often called the method of scoring. The main feature of the scheme (3.1.6) is that t is fixed, and ϑ_k, at each step $k = 1, 2, \ldots$, is the k'th approximation to a root, say

$\tilde{\theta}_t$, of the likelihood equation $\sum_{s=1}^{t} (f'(X_s, v)/f(X_s, v)) = 0$. Also, if a new $(t+1)$st observation is available, the whole procedure has to be repeated again. Note also, that the one-step Newton-Raphson is a simplified version of (3.1.6) when an auxiliary $\sqrt{t}$-consistent estimator, say $\tilde{\theta}_t$ is available. Then, the one-step Newton-Raphson improves $\tilde{\theta}_t$ in one step (that is, $k = 1$) by

$$\hat{\theta}_t = \tilde{\theta}_t + J^{-1}(\tilde{\theta}_t) \sum_{s=1}^{t} \frac{f'(X_s, \tilde{\theta}_t)}{f(X_s, \tilde{\theta}_t)}. \tag{3.1.7}$$

As one can see the procedure (3.1.3) is quite different. It does not require an auxiliary estimator and it adjusts the value of the estimator at each instant of time with the arrival of the new observation. A theoretical implication of this is that by studying the procedures (3.1.3), or in general (3.1.4), we study the asymptotic behaviour of the estimator. As far as applications are concerned, there are advantages in using (3.1.3), (3.1.4), or (3.1.5), since these procedures are easy to use and, unlike other methods, do not require storing all the data. Also, these procedures naturally allow for on-line implementation, which is particulary convenient for sequential data processing.

It should be noted that the recursive procedure (3.1.4) is not a numerical solution of (3.1.1). Nevertheless, recursive estimator (3.1.4) and the corresponding M-estimator are expected to have the same asymptotic properties under quite mild conditions.

As it was mentioned in Section 1.1, in the i.i.d. case, (3.1.3) can be regarded as a classical stochastic approximation procedure and in the general statistics model,

(3.1.4) can be rewritten in the SA form by introducing

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

Following the argument in Remark 2.4.4 (see also Sharia [69]), an optimal step-size sequence would be

$$\gamma_t^{-1}(v) = -\sum_{s=1}^{t} R_s'(v) \tag{3.1.8}$$

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

$$\gamma_t(v)^{-1} = -\sum_{s=1}^{t} E_\theta\{\psi_s'(v) \mid \mathcal{F}_{s-1}\}, \tag{3.1.9}$$

or any sequence with the increments

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

Also, since $\psi_t(\theta)$ is 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 [69] for detials)

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

where $l_t(\theta) = [f'_t(\theta, X_t|X_1^{t-1})]^T / f_t(\theta, X_t|X_1^{t-1})$. Therefore, another possible choice of the normalizing 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 the mutual quadratic characteristic of the martingales

$$\gamma_t^{-1}(\theta) = \sum_{s=1}^{t} E_\theta\{\psi_s(\theta) l_s^T(\theta)|\mathcal{F}_{t-1}\} = \langle M^\theta, L^\theta\rangle_t$$

where $L_t^\theta = \sum_{s=1}^t l_s(\theta)$ 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 $I_t(\theta)$, and the corresponding recursive procedure is

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} + I_t^{-1}(\hat{\theta}_{t-1}) l_t(\hat{\theta}_{t-1})\Big), \qquad t \geq 1, \tag{3.1.10}$$

Also, given that the model possesses certain ergodicity properties, asymptotic linearity of (3.1.10) 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 $(0,\ i^{-1}(\theta))$.

3.2 Special models and examples

3.2.1 The i.i.d. case

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

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

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

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} + a_t^{-1}\gamma(\hat{\theta}_{t-1})\psi(\hat{\theta}_{t-1}, X_t)\Big) \tag{3.2.1}$$

where a_t is a non-decreasing real sequence, $\gamma(\theta)$ is a invertible $m \times m$ matrix and truncation sequence U_t is admissible for θ.

The i.i.d. scheme can be analysed in the framework of classical stochastic approximation. Everywhere in the following example, $\mathcal{F}_t$ is the σ-algebra generated by $X_1, \ldots, X_t$, P^θ is the family of corresponding measures, and $\theta > 0$ is an arbitrary but fixed value of the parameter. It is easy to see that (3.2.1) can be rewritten in the form of (2.5.1) with

$$\begin{aligned} Z_t &= \hat{\theta}_t \\ \gamma_t(u) &= a_t^{-1}\mathbf{I} \\ R(u) &= \gamma(u) E_\theta \{\psi(u, X_t)\} \\ \varepsilon_t(u) &= \gamma(u)\psi(u, X_t) - R(u) \end{aligned}$$

Example 3.2.1 Let $X_1, X_2, \ldots$ be i.i.d. random variables from Gamma$(\theta, 1)$ $(\theta > 0)$. Then 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 that

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

Then

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

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

$$\hat{\theta}_t = \Phi_{U_t}\left(\hat{\theta}_{t-1} + \frac{1}{t\,\log''\mathbf{\Gamma}(\hat{\theta}_{t-1})}\left[\log X_t - \log'\mathbf{\Gamma}(\hat{\theta}_{t-1})\right]\right) \tag{3.2.2}$$

with $U_t = [\alpha_t, \beta_t]$ where $\alpha_t \downarrow 0$ and $\beta_t \uparrow \infty$ are sequences of positive numbers.

Let us rewrite (3.2.2) in the form of the stochastic approximation, i.e.,

$$\hat{\theta}_t = \Phi_{U_t}\left(\hat{\theta}_{t-1} + \frac{1}{t}\left[R(\hat{\theta}_{t-1}) + \varepsilon_t(\hat{\theta}_{t-1})\right]\right) \tag{3.2.3}$$

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

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

Since $E_\theta \{\log X_t \mid \mathcal{F}_{t-1}\} = E_\theta \{\log X_t\} = \log' \mathbf{\Gamma}(\theta)$ and $\hat{\theta}_{t-1}$ is $\mathcal{F}_{t-1}$ - measurable, we have $E_\theta \left\{\varepsilon_t(\hat{\theta}_{t-1}) \mid \mathcal{F}_{t-1}\right\} = 0$ and hence (2.1.2) holds. Since $E_\theta \left\{\log^2 X_t\right\} < \infty$, condition (2.1.3) can be checked in the similar way. Obviously, $R(\theta) = 0$, and since $\log' \mathbf{\Gamma}$ is increasing (see, e.g., Whittaker and Watson [76], 12.16), condition (H1) of Corollary 2.5.1 and conditions of Corollary 2.5.5 hold with $z^0 = \theta$ and $a_t = t$ (see Appendix B for details). Then it can be shown that if

$$\sum_{t=1}^{\infty} \frac{\alpha_{t-1}^2}{t} = \infty \qquad \text{and} \qquad \sum_{t=1}^{\infty} \frac{\log^2 \alpha_{t-1} + \log^2 \beta_{t-1}}{t^2} < \infty, \tag{3.2.4}$$

then all the conditions of Corollary 2.5.1 hold and therefore, $\hat{\theta}_t$ is strongly consistent, i.e.,

$$\hat{\theta}_t \xrightarrow{a.s.} \theta \qquad \text{as} \qquad t \longrightarrow \infty.$$

For instance, the sequences,

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

with some positive constants C_1 and C_2, obviously satisfy (3.2.4).

Note that since $\theta \in (0, \infty)$, it may seem unnecessary to use the upper truncations $\beta_t < \infty$. However, without upper truncations (i.e. if $\beta_t = \infty$), the standard restriction on the growth of R does not hold. Also, with $\beta_t = \infty$ the procedure fails condition (H2) of Corollary 2.5.1.

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 2.5.7 holds. Since $E_\theta\left\{\varepsilon_t(u) \mid \mathcal{F}_{t-1}\right\} = 0$, we have

$$E_\theta\left\{[R(u) + \varepsilon(u)]^2 \mid \mathcal{F}_{t-1}\right\} = R^2(u) + E_\theta\left\{\varepsilon_t^2(u) \mid \mathcal{F}_{t-1}\right\}. \tag{3.2.5}$$

Using (3.2.5) and (B.5) in Appendix B,

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

which is obviously a continuous function of u.

Thus, for any $v_t \longrightarrow 0$, $E_\theta\left\{\varepsilon_t^2(\theta + v_t) \mid \mathcal{F}_{t-1}\right\}$ converges to a finite limit and so condition (BB) in Corollary 2.5.9 holds. Therefore, conditions in Corollary 2.5.9 are satisfied with $a_t = t$ and we have $t^\delta(\hat{\theta}_t - \theta)^2 \xrightarrow{a.s.} 0$ 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 2.5.11

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 (3.2.3) is asymptotic linear. Now, using the CLT for i.i.d. r.v.'s, it follows that $\hat{\theta}_t$ is asymptotically efficient.

Remark 3.2.2 Note that the discussion of convergence in the above example is from Sharia [70].

3.2.2 Exponential family of Markov processes

Consider a time-homogeneous Markov process X_t (t=1,2...). We say that X_t belongs to the exponential family of Markov processes, if the conditional probability density function of X_t given X_{t-1} is $f_t(\theta, x_t|x_{t-1}) = f(x_t;\theta, x_{t-1})$, where

$$f(y;\theta,x) = h(x,y)exp\{\theta^T m(y,x) - \beta(\theta;x)\}$$

where $m(y,x)$ is a $1\times m$ vector and $\beta(\theta;x)$ is a scalar.

It follows from the standard exponential family theory (see Feigin [24]) that

$$l_t(\theta) = \frac{d}{d\theta}logf(X_t;\theta,X_{t-1}) = m(X_t,X_{t-1}) - [\beta'(\theta;X_{t-1})]^T$$

is a martingale-difference and the conditional Fisher information is

$$I_t(\theta) = \sum_{s=1}^{t}\beta''(\theta;X_{s-1}).$$

Therefore, a maximum likelihood recursive procedure can be defined as

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \Big(\sum_{s=1}^{t} \beta''(\hat{\theta}_{s-1}; X_{s-1}) \Big)^{-1} \Big(m(X_t, X_{t-1}) - [\beta'(\hat{\theta}_{t-1}; X_{t-1})]^T \Big)$$

for $t \geq 1$.

Locating the actual value of parameter θ is now the same as finding the root of function $R_t(z) = E_\theta\{l_t(z)|\mathcal{F}_{t-1}\}$.

Since $E_\theta\{l_t(\theta)|\mathcal{F}_{t-1}\} = 0$, we have

$$E_\theta\{m(X_t, X_{t-1})|\mathcal{F}_{t-1}\} = [\beta'(\theta; X_{t-1})]^T$$

and

$$E_\theta\{l_t(\theta + u)\} = [\beta'(\theta; X_{t-1}) - \beta'(\theta + u; X_{t-1})]^T.$$

Now suppose that θ is one-dimensional and consider the class of conditionally additive exponential families, that is,

$$f(y; \theta, x) = h(x, y) exp\Big\{ \theta m(y, x) - \varphi(\theta) h(x) \Big\},$$

where $h(\cdot) \geq 0$ and $\varphi''(\cdot) \geq 0$. Then the conditional Fisher information is

$$I_t(\theta) = \varphi''(\theta) H_t \quad \text{where} \quad H_t = \sum_{s=1}^{t} h(X_{s-1}).$$

and

$$E_\theta\{l_t^2(\theta)|\mathcal{F}_{t-1}\} = \varphi''(\theta) h(X_{t-1}) \tag{3.2.6}$$

(see details also in Feigin [24]).

Therefore, a maximum likelihood recursive procedure can be defined as

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \Big(\varphi''(\hat{\theta}_{t-1})H_t\Big)^{-1}\Big(m(X_t, X_{t-1}) - \varphi'(\hat{\theta}_{t-1})h(X_{t-1})\Big) \tag{3.2.7}$$

for $t \geq 1$.

We can rewrite (3.2.7) in the form of (2.1.1) with

$$\begin{aligned} Z_t &= \hat{\theta}_t \\ \gamma_t(\theta + u) &= \Big(\varphi''(\theta + u)H_t\Big)^{-1} \\ R_t(\theta + u) &= [\varphi'(\theta) - \varphi'(\theta + u)]h(X_{t-1}) \\ \varepsilon_t(\theta + u) &= l_t(\theta + u) - E_\theta\{l_t(\theta + u)|\mathcal{F}_{t-1}\}, \end{aligned}$$

and truncations $U_t = \mathbb{R}^m$.

Lemma 3.2.3 *Let $\hat{\theta}_t$ be estimators defined by* (3.2.7). *Suppose that $H_t \xrightarrow{a.s.} \infty$ and either φ' is a linear function or the following conditions hold:*

(M1)

$$\frac{h(X_{t-1})}{H_t} \xrightarrow{a.s.} 0;$$

(M2) *for any finite a and b,*

$$0 < \inf_{u\in[a,b]} \varphi''(u) \leq \sup_{u\in[a,b]} \varphi''(u) < \infty;$$

(M3) *there exists a constant K such that*

$$\frac{1 + [\varphi'(u)]^2}{[\varphi''(u)]^2} \leq K(1 + u^2)$$

for each $u \in \mathbb{R}$.

Then $\hat{\theta}_t$ *converges to* θ *(a.s.), for any initial value* $\hat{\theta}_0$.

Proof. See details in Sharia [66].

Lemma 3.2.4 *Suppose that* $\hat{\theta}_t \xrightarrow{a.s.} \theta$ *where* $\hat{\theta}_t$ *is defined by* (3.2.7) *with* $H_t \longrightarrow \infty$. *Suppose also that* $\varphi''(\cdot)$ *is a continuous function and is positive in a neighbourhood of* θ. *Then* $H_t^\delta(\hat{\theta}_t - \theta)^2 \xrightarrow{a.s.} 0$ *for any* $\delta < 1$.

Proof. Let us check the conditions of Lemma 2.3.7 with $C_t = H_t^\delta \mathbf{I}$, $\rho_t = h(X_{t-1})/H_t$ and $\mathcal{P}_t = 0$. Since $U_t = \mathbb{R}$, condition (V1) of Lemma 2.2.1 holds trivially.

Since

$$\varphi'(\theta) - \varphi'(\theta + u) = -u\varphi''(\theta + \tilde{u}) \quad \text{with} \quad |\tilde{u}| \le |u|,$$

we have

$$\begin{aligned} & \left[\frac{2\Delta_{t-1} C_t \gamma_t(\theta + \Delta_{t-1}) R_t(\theta + \Delta_{t-1})}{\lambda_t^{max}} + \mathcal{P}_t\right] \frac{1}{\rho_t \Delta_{t-1}^2} \\ = & \left[-2\frac{H_t^\delta h(X_{t-1})\varphi''(\theta + \tilde{\Delta}_{t-1})}{H_t \varphi''(\theta + \Delta_{t-1})}\Delta_{t-1}^2\right] \frac{H_t}{h(X_{t-1})\Delta_{t-1}^2} \\ = & -2\frac{\varphi''(\theta + \tilde{\Delta}_{t-1})}{\varphi''(\theta + \Delta_{t-1})} H_t^\delta \end{aligned}$$

Since $\Delta_t \longrightarrow 0$, we have $\tilde{\Delta}_t \longrightarrow 0$ and $= \varphi''(\theta + \tilde{\Delta}_{t-1})/\varphi''(\theta + \Delta_{t-1}) \longrightarrow 1$. Therefore,

$$\left[\frac{2\Delta_{t-1}^T C_t \gamma_t(Z_{t-1}) R_t(Z_{t-1})}{\lambda_t^{max}} + \mathcal{P}_t\right] \frac{1}{\rho_t \Delta_{t-1}^2} < -1$$

eventually. Meantime, since $h_t \ge 0$ and $(H_t/H_{t-1})^\delta \le H_t/H_{t-1}$ (indeed, $H_t \ge$

$H_{t-1} > 0$ and $\delta \leq 1$) and $h(X_{t-1})/H_t < 1$, we have

$$\begin{aligned}
& \sum_{t=1}^{\infty} \left[\frac{\lambda_t^{max} - \lambda_{t-1}^{min}}{\lambda_{t-1}^{min}} - \frac{\lambda_t^{max}}{\lambda_{t-1}^{min}} \rho_t \right]^+ \\
= & \sum_{t=1}^{\infty} \left[\frac{H_t^\delta - H_{t-1}^\delta}{H_{t-1}^\delta} - \frac{H_t^\delta}{H_{t-1}^\delta} \frac{h(X_{t-1})}{H_t} \right]^+ \\
= & \sum_{t=1}^{\infty} \left[(1 - \frac{h(X_{t-1})}{H_t})(\frac{H_t}{H_{t-1}})^\delta - 1 \right]^+ \\
\leq & \sum_{t=1}^{\infty} \left[(1 - \frac{h(X_{t-1})}{H_t}) \frac{H_t}{H_{t-1}} - 1 \right]^+ \\
= & \sum_{t=1}^{\infty} \left[\frac{H_t - h(X_{t-1})}{H_{t-1}} - 1 \right]^+ \\
= & 0 < \infty.
\end{aligned}$$

Condition (R1) has been satisfied.

According to Proposition A.8 in Appendix A, $\sum_{t=1}^{\infty} h(X_{t-1})/H_t^{2-\delta} < \infty$, and by (3.2.6),

$$\begin{aligned}
& \sum_{t=1}^{\infty} \frac{\lambda_t^{max} \left[E \left\{ \|\gamma_t(Z_{t-1})[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})]\|^2 \mid \mathcal{F}_{t-1} \right\} - \mathcal{P}_t \right]^+}{1 + \lambda_{t-1}^{min} \|\Delta_{t-1}\|^2} \\
\leq & \sum_{t=1}^{\infty} \lambda_t^{max} \left[\gamma_t^2(\theta + \Delta_{t-1}) E \left\{ l_t^2(\theta + \Delta_{t-1}) \mid \mathcal{F}_{t-1} \right\} \right]^+ \\
\leq & \sum_{t=1}^{\infty} \left[\frac{h(X_{t-1})}{\varphi''(\theta + \Delta_{t-1}) H_t^{2-\delta}} \right]^+ \\
< & \infty
\end{aligned}$$

as $\Delta_t \longrightarrow 0$. (R2) is now satisfied. Therefore, by Lemma 2.3.7, $H_t^\delta(\hat{\theta}_t - \theta)^2 \xrightarrow{a.s.} 0$.■

Lemma 3.2.5 *Suppose that condition M*1 *and M*3 *hold in Lemma 3.2.3. Then* $\hat{\theta}_t$ *is asymptotically linear, if function* $\varphi''(\cdot)$ *is locally Lipschitz, that is, for any* θ *there*

exists a constant K_θ *and* $0 < \varepsilon_\theta \leq 1/2$ *such that*

$$|\varphi''(\theta + u) - \varphi''(\theta)| \leq K_\theta |u|^{\varepsilon_\theta}.$$

Proof. See details in Sharia [69].

3.2.3 Linear procedures

Consider the linear recursive procedure

$$Z_t = Z_{t-1} + \gamma_t(h_t - \beta_t Z_{t-1}) \tag{3.2.8}$$

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 $t \geq 1$). The following result gives a sufficient condition for convergence and asymptotic linearity of the estimator defined by (3.2.8) in the case when $h_t - \beta_t z^0$ is a martingale-difference, i.e., $E\{h_t|\mathcal{F}_{t-1}\} = \beta_t z^0$. We can view (3.2.8) as a SA procedure designed for finding z^0, which is the root of function

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

with the random noise

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

Lemma 3.2.6 *Suppose that* Z_t *is defined by* (3.2.8)*,* C_t *is a sequence of* $m \times m$ *positive semi-definite matrices and*

(J1) $\Delta C_t - 2C_t\gamma_t\beta_t + \beta_t\gamma_t C_t\gamma_t\beta_t$ *is negative semi-definite eventually;*

(J2)

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

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

proof. Consider Lemma 2.2.1 with $V_t(u) = u^T C_t u$ and truncations $U_t = \mathbb{R}^m$, then (V1) holds trivially and we have $V_t'(u) = 2u^T C_t$ and $V_t''(u) = 2C_t$. Then

$$\begin{aligned} V_t'(\Delta_{t-1})\gamma_t(z^0 + \Delta_{t-1})R_t(z^0 + \Delta_{t-1}) &= 2\Delta_{t-1}^T C_t\gamma_t(Z_{t-1})R_t(Z_{t-1}) \\ &= -\Delta_{t-1}^T 2C_t\gamma_t\beta_t\Delta_{t-1} \ . \end{aligned}$$

Since ε_t is a martingale-difference,

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

Hence,

$$
\begin{aligned}
& \sum_{t=1}^{\infty} \frac{[\mathcal{K}_t(\Delta_{t-1})]^+}{1+V_{t-1}(\Delta_{t-1})} \\
\leq & \sum_{t=1}^{\infty} \Big[\Delta V_t(\Delta_{t-1}) + V_t'(\Delta_{t-1})\gamma_t R_t(Z_{t-1}) \\
& + \frac{1}{2} E\Big\{ \big[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})\big]^T \gamma_t(Z_{t-1}) V_t'' \gamma_t(Z_{t-1}) \big[R_t(Z_{t-1}) + \varepsilon_t(Z_{t-1})\big] | \mathcal{F}_{t-1} \Big\} \Big]^+ \\
= & \sum_{t=1}^{\infty} \Big[\Delta_{t-1}^T (\Delta C_t - 2C_t\gamma_t\beta_t + \beta_t\gamma_t C_t\gamma_t\beta_t)\Delta_{t-1} \\
& + E\Big\{ (h_t - \beta_t z^0)^T \gamma_t C_t \gamma_t (h_t - \beta_t z^0) | \mathcal{F}_{t-1} \Big\} \Big]^+ \\
\leq & \sum_{t=1}^{\infty} \Big([\Delta_{t-1}^T (\Delta C_t - 2C_t\gamma_t\beta_t + \beta_t\gamma_t C_t\gamma_t\beta_t)\Delta_{t-1}]^+ \\
& + E\Big\{ (h_t - \beta_t z^0)^T \gamma_t C_t \gamma_t (h_t - \beta_t z^0) | \mathcal{F}_{t-1} \Big\} \Big).
\end{aligned}
$$

Thus, by (J1) and (J2),

$$
\sum_{t=1}^{\infty} \frac{[\mathcal{K}_t(\Delta_{t-1})]^+}{1+V_{t-1}(\Delta_{t-1})} \leq E\Big\{ (h_t - \beta_t z^0)^T \gamma_t C_t \gamma_t (h_t - \beta_t z^0) | \mathcal{F}_{t-1} \Big\} < \infty.
$$

Therefore, condition (V2) holds. According to Lemma 2.2.1, $(Z_t - z^0)^T C_t (Z_t - z^0)$ converges to a finite limit (a.s.).

■

Corollary 3.2.7 *Suppose that Z_t is defined by* (3.2.8)*, a_t is an 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 (a.s.).*

Proof. Consider Lemma 3.2.6 with $C_t = (a_t\gamma_t)^{-1}$, condition (J2) is satisfied immediately.

For all $u \in \mathbb{R}^m$, since

$$\begin{aligned} & u^T \Delta C_t u = u^T[(a_t\gamma_t)^{-1} - (a_{t-1}\gamma_{t-1})^{-1}]u \\ \leq \;& u^T(a_t\gamma_t)^{-1}u - u^T(a_t\gamma_{t-1})^{-1}u = u^T a_t^{-1}\Delta\gamma_t^{-1}u, \end{aligned}$$

we have

$$u^T(\Delta C_t - 2C_t\gamma_t\beta_t + \beta_t\gamma_t C_t\gamma_t\beta_t)u \leq u^T a_t^{-1}(\Delta\gamma_t^{-1} - 2\beta_t + \beta_t\gamma_t\beta_t)u$$

is non-negative eventually. Condition (J1) is now satisfied. Therefore, $a_t^{-1}(Z_t - z^0)^T\gamma_t^{-1}(Z_t - z^0)$ converges to a finite limit (a.s.). ■

Remark 3.2.8 In the case when the minimum eigenvalue λ_t^{min} of C_t goes to infinity, it can be derived that $\|Z_t - z^0\| \longrightarrow 0$ (a.s.) if all conditions hold in Lemma 3.2.6.

Corollary 3.2.9 *Suppose that* $\Delta\gamma_t^{-1} = \beta_t$*, then (G1) in Corollary 3.2.7 holds.*

Proof. Since $\Delta\gamma_t^{-1}$ is positive semi-definite, $\Delta\gamma_t$ is negative semi-definite $\Big($see Horn and Johnson [32], Corollary 7.7.4(a)$\Big)$ and

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

is also negative semi-definite. ■

Proposition 3.2.10 *Suppose that Z_t is defined by* (3.2.8), $\gamma_t \longrightarrow 0$ *and*

$$\gamma_t^{1/2}\sum_{s=1}^{t}(\Delta\gamma_s^{-1} - \beta_s)\Delta_{s-1} \longrightarrow 0 \tag{3.2.9}$$

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

Then Z_t is asymptotically linear, that is,

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

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

Proof. Let us check the conditions of Theorem 2.4.1 for $A_t = \gamma_t^{-1/2}$. Condition (E1) and (E2) trivially holds. Since $\varepsilon_t(u) = h_t - \beta_t z^0$ is state free, (E4) also holds.

Since β_t and Z_{t-1} are $\mathcal{F}_{t-1}$-measurable,

$$\begin{aligned}\tilde{R}_t(Z_{t-1}) = R_t(Z_{t-1}) &= E\{h_t - \beta_t Z_{t-1}|\mathcal{F}_{t-1}\} \\ &= E\{h_t|\mathcal{F}_{t-1}\} - \beta_t Z_{t-1} \\ &= \beta_t z^0 - \beta_t Z_{t-1} \\ &= -\beta_t \Delta_{t-1} \quad .\end{aligned}$$

Therefore,

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

and (E3) is equivalent to (3.2.9). Thus, all the conditions of Theorem 2.4.1 hold, and Z_t is asymptotically linear. ■

3.3 Summary

This chapter illustrates how the SA theory is applied to recursive parameter estimation for the general statistical model. As it is mentioned in Section 3.1, the recursive estimation procedure given in this chapter is not an approximate solution of the corresponding estimating equation. However, it shares the same asymptotic properties as the corresponding non-recursive estimator. In addition, the recursive procedures do not require storing all the data to generate new estimates. These results demonstrate that the SA type recursive likelihood estimation requires minimum assumptions for the i.i.d. model (see Section 3.2). The result on rate of convergence in Section 3.2.2 improves the previous results in recursive estimation by Sharia [66] and [69], derived in the context of statistical parametric estimation.

A new set of conditions to derive the rate of convergence for Linear processes is also derived in Section 3.2.3.

Chapter 4

Parameter Estimation in Autoregressive Models

In this chapter, we use the results obtained in the previous chapters to study on-line procedures for AR(m) processes.

In Section 4.1 we propose a general class of on-line recursive procedures using the ideas introduced in Chapter 3. In Section 4.2 we study asymptotic behaviour of the recursive LS estimators. In Section 4.3 we present asymptotic results for the general class of truncated recursive estimators, which includes recursive MLE procedures. An example of an AR process with the Students innovations is also presented to demonstrate how the SA theory works for AR models.

4.1 On-line recursive estimators

Consider an AR(m) process

$$X_t = \theta^{(1)} X_{t-1} + \theta^{(2)} X_{t-2} + \cdots + \theta^{(m)} X_{t-m} + \xi_t = \theta^T X_{t-m}^{t-1} + \xi_t \qquad (4.1.1)$$

where $\theta = (\theta^{(1)}, ..., \theta^{(m)})^T$, $X_{t-m}^{t-1} = (X_{t-1}, ..., X_{t-m})^T$ and ξ_t is a sequence of independent random variables. If the pdf of ξ_t w.r.t. Lebesgue's measure is $g_t(x)$, then the conditional distribution function of X_t given the past observation 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

$$\frac{f'^T_t(\theta, x | X_1^{t-1})}{f_t(\theta, x | X_1^{t-1})} = -\frac{g'_t(x - \theta^T X_{t-m}^{t-1})}{g_t(x - \theta^T X_{t-m}^{t-1})} X_{t-m}^{t-1}. \tag{4.1.2}$$

Also, the one-step conditional Fisher information at t is

$$\begin{aligned}
& E_\theta \left\{ \frac{f'^T_t(\theta, X_t) f'_t(\theta, X_t)}{f_t^2(\theta, X_t)} \Big| \mathcal{F}_{t-1} \right\} \\
= \; & E_\theta \left\{ \left[\frac{g'_t(X_t - \theta^T X_{t-m}^{t-1})}{g_t(X_t - \theta^T X_{t-m}^{t-1})} \right]^2 X_{t-m}^{t-1} (X_{t-m}^{t-1})^T \Big| \mathcal{F}_{t-1} \right\} \\
= \; & X_{t-m}^{t-1} (X_{t-m}^{t-1})^T E_\theta \left\{ \left[\frac{g'_t(X_t - \theta^T X_{t-m}^{t-1})}{g_t(X_t - \theta^T X_{t-m}^{t-1})} \right]^2 \Big| \mathcal{F}_{t-1} \right\} \\
= \; & X_{t-m}^{t-1} (X_{t-m}^{t-1})^T \int_{-\infty}^{\infty} \left[\frac{g'_t(x - \theta^T X_{t-m}^{t-1})}{g_t(x - \theta^T X_{t-m}^{t-1})} \right]^2 g_t(x - \theta^T X_{t-m}^{t-1}) dx \\
= \; & X_{t-m}^{t-1} (X_{t-m}^{t-1})^T l_{gt}
\end{aligned}$$

where

$$l_{gt} = \int_{-\infty}^{+\infty} \left(\frac{g'_t(x)}{g_t(x)} \right)^2 g_t(x) dx.$$

Therefore, the conditional Fisher information matrix is now

$$I_t = \sum_{s=1}^{t} l_{gs} X_{s-m}^{s-1} (X_{s-m}^{s-1})^T.$$

Note also that I_t^{-1} can be generated as

$$I_t^{-1} = I_{t-1}^{-1} - l_{gt} I_{t-1}^{-1} X_{t-m}^{t-1} (1 + l_{gt} (X_{t-m}^{t-1})^T I_{t-1}^{-1} X_{t-m}^{t-1})^{-1} (X_{t-m}^{t-1})^T I_{t-1}^{-1}. \tag{4.1.3}$$

(Recursion (4.1.3) is known as the Riccati equation. See also Lemma A.5 in Appendix A for a simple proof.)

So, the on-line likelihood procedure introduced in Section 3.1 in this case can be derived by the following recursion

$$\hat{\theta}_t = \hat{\theta}_{t-1} - I_t^{-1} X_{t-m}^{t-1} \frac{g_t'}{g_t} (X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1}) \tag{4.1.4}$$

where I_t^{-1} is also derived on-line using formula (4.1.3).

In general, to include robust estimation procedures, and also to use any available auxiliary information, one can use the following class of procedures

$$\hat{\theta}_t = \Phi_{U_t} \left(\hat{\theta}_{t-1} + \gamma_t H(X_{t-m}^{t-1}) \varphi_t (X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1}) \right), \tag{4.1.5}$$

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 a $m \times m$ matrix valued step-size sequence.

Here, truncation sequence U_t represents auxiliary knowledge about the unknown parameter which is incorporated in the procedure through the truncation operator Φ. For example, for an $AR(2)$ process, if the roots of the corresponding polynomial lie outside of the unit circle, one can take $U_t = U$ where U is a triangle defined by

$$U = \{ \ (\theta^{(1)}, \theta^{(2)}) \ : \ |\theta^{(2)}| < 1, \ \theta^{(1)} + \theta^{(2)} < 1, \ \theta^{(2)} - \theta^{(1)} < 1 \ \}. \tag{4.1.6}$$

As we can see in Section 4.2, for AR processes one can always construct on-line LS estimators, $\hat{\theta}_t^{LS}$, which are consistent under very mild conditions. However, the LS estimators are not asymptotically efficient unless the innovations are Gaussian r.v.'s. Therefore, one can construct the following system of on-line procedures using the LS truncations

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} - I_t^{-1} X_{t-m}^{t-1} \frac{g_t'}{g_t}(X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1})\Big) \tag{4.1.7}$$

where I_t^{-1} is derived by (4.1.3) and $U_t = U_t(\hat{\theta}_t^{LS})$ is sequence of sets that converge to the "true" value of the parameter θ. For example, one of the possible choices for U_t is a sphere in $\mathbb{R}^m$ with the center at $\hat{\theta}_t^{LS}$ and radius $ct^{-\varepsilon}$, where c is a positive constant and $0 < \varepsilon < 1/2$ (see Example 4.3.6).

4.2 Recursive least squares procedures

Recursive least squares (RLS) estimator of $\theta = (\theta^{(1)}, \ldots, \theta^{(m)})^T$ is generated (see Lai and Wei [44]) by the following procedure

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \hat{I}_t^{-1} X_{t-m}^{t-1}[X_t - (X_{t-m}^{t-1})^T \hat{\theta}_{t-1}], \tag{4.2.1}$$

$$\hat{I}_t^{-1} = \hat{I}_{t-1}^{-1} - \hat{I}_{t-1}^{-1} X_{t-m}^{t-1}[1 + (X_{t-m}^{t-1})^T \hat{I}_{t-1}^{-1} X_{t-m}^{t-1}]^{-1} (X_{t-m}^{t-1})^T \hat{I}_{t-1}^{-1}. \tag{4.2.2}$$

It is easy to see that (4.2.1) is a special case of the linear process (3.2.8) with

$$\begin{aligned}\gamma_t &= \hat{I}_t^{-1} \quad (=\gamma_t^T),\\ h_t &= X_{t-m}^{t-1}X_t = X_{t-m}^{t-1}\Big[(X_{t-m}^{t-1})^T z^0 + \xi_t\Big],\\ \beta_t &= X_{t-m}^{t-1}(X_{t-m}^{t-1})^T.\end{aligned}$$

Corollary 4.2.1 *Consider $\hat{\theta}_t$ defined by* (4.2.1) *and* (4.2.2)*. If 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 (a.s.).

Proof. Check the condition of Corollary 3.2.7. Obviously, matrix $\gamma_t = \hat{I}_t^{-1} = c\cdot\mathbf{I} + \sum_{s=1}^{t} X_{s-m}^{s-1}(X_{s-m}^{s-1})^T$ 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 3.2.9, condition (G1) holds. We also have

$$\begin{aligned}&\sum_{t=1}^{\infty} a_t^{-1} E\{(h_t - \beta_t z^0)^T \gamma_t (h_t - \beta_t z^0)|\mathcal{F}_{t-1}\}\\ =\ &\sum_{t=1}^{\infty} a_t^{-1} E\{\xi_t (X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1}\xi_t|\mathcal{F}_{t-1}\}\\ =\ &\sum_{t=1}^{\infty} a_t^{-1}(X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} E\{\xi_t^2|\mathcal{F}_{t-1}\}\\ <\ &\infty,\end{aligned}$$

which leads to (G2). Therefore, conditions of Corollary 3.2.7 hold. ■

Corollary 4.2.2 *Consider $\hat{\theta}_t$ defined by* (4.2.1) *and* (4.2.2)*. Suppose that*

(P1) *There exists a non-decreasing sequence* $\{\kappa_t\}$ *such that*

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

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

(P2) $\{\xi_t\}$ *is a martingale-difference and*

$$E\left\{\xi_t^2 \mid \mathcal{F}_{t-1}\right\} = O(\kappa_t^{\delta})$$

(a.s.) for any $\delta > 0$.

Then $\kappa_t^{1-\delta}\|\hat{\theta}_t - \theta\|^2 \longrightarrow 0$ *(a.s.) for all* $\delta > 0$.

Proof. Concider Corollary 4.2.1 with $a_t = \kappa_t^{\delta}$ for a certain $\delta > 0$. Since $E\{\xi_t^2 \mid \mathcal{F}_{t-1}\} = O(\kappa_t^{\delta/2})$, there exists (a possibly positive) constant K such that

$$\begin{aligned}\sum_{t=1}^{\infty} a_t^{-1}(X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} E\{\xi_t^2|\mathcal{F}_{t-1}\} &\leq \sum_{t=1}^{\infty} \kappa_t^{-\delta}(X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} K \kappa_t^{\delta/2} \\ &= K \sum_{t=1}^{\infty} \kappa_t^{-\delta/2}(X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1}\end{aligned}$$

According to Lemma A.3 in the Appendix A,

$$\sum_{t=1}^{\infty} \kappa_t^{-\delta/2}(X_{t-m}^{t-1})^T \hat{I}_t^{-1} X_{t-m}^{t-1} < \infty$$

Therefore, $(\hat{\theta}_t - \theta)^T \hat{I}_t(\hat{\theta}_t - \theta)/\kappa_t^{\delta}$ tends to a finite limit for all $\delta > 0$. As $\hat{I}_t/\kappa_t$ tends to a finite matrix, it can be deduced that $\kappa_t^{1-\delta}\|\hat{\theta}_t - \theta\|^2$ tends to a finite limit. ■

Remark 4.2.3 If X_t defined by (4.1.1) is a strongly stationary process, then one can take $\kappa_t = t$ and $t^{1-\delta}\|\hat{\theta}_t - \theta\|^2$ converges to a finite limit by Corollary 4.2.2.

Note that the convergence of the LS estimator is well known under the stationary assumption. (see e.g., Shiryayev [71], Ch.VII, §5). This section is presented to demonstrate that the conditions made here are minimal. That is, for well-known models, the results of this work do not assume any additional restrictions. Moreover, convergence can be derived using the results given above, without the stationary 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.

4.3 On-line recursive M-estimators with truncations

Consider an AR(m) process, X_t, defined by (4.1.1) and denote by $g_t(x)$ the pdf of ξ_t w.r.t. Lebesgue's measure. For all $v \in \mathbb{R}$, define

$$P_t(v) = \int_{-\infty}^{\infty} \varphi_t(z - v)\, g_t(z)dz$$

and

$$Q_t(v) = \int_{-\infty}^{\infty} [\varphi_t(z - v)]^2\, g_t(z)dz.$$

Recursive estimating procedure (4.1.5) can be considered in the form of (2.1.1) with

$$\begin{aligned} Z_t &= \hat{\theta}_t \\ R_t(u) &= H(X_{t-m}^{t-1})E_\theta\left\{\varphi_t(X_t - u^T X_{t-m}^{t-1})\Big|\mathcal{F}_{t-1}\right\} \\ &= H(X_{t-m}^{t-1})E_\theta\left\{\varphi_t[\xi_t - (u-\theta)^T X_{t-m}^{t-1}]\Big|\mathcal{F}_{t-1}\right\} \\ &= H(X_{t-m}^{t-1})P_t\Big((u-\theta)^T X_{t-m}^{t-1}\Big) \\ \varepsilon_t(u) &= H(X_{t-m}^{t-1})\varphi_t(X_t - u^T X_{t-m}^{t-1}) - R_t(u) \\ &= H(X_{t-m}^{t-1})\varphi_t\Big(\xi_t + (\theta-u)^T X_{t-m}^{t-1}\Big) - R_t(u). \end{aligned} \tag{4.3.1}$$

Corollary 4.3.1 *Suppose that $\hat{\theta}_t$ is generated by* (4.1.5)*, condition (V1) of Lemma 2.2.1 holds and there exists a non-decreasing predictable process $a_t > 0$ such that*

(T1) *for all $u \in \mathbb{R}^{m\times 1} \cap U_{t-1}$,*

$$u^T(a_t^{-1}\gamma_t^{-1} - a_{t-1}^{-1}\gamma_{t-1}^{-1})u + 2a_t^{-1}u^T H(X_{t-m}^{t-1})P_t(u^T X_{t-m}^{t-1}) \le 0$$

eventually;

(T2) *for any predictable vector process $d_t \in U_{t-1}$,*

$$\sum_{t=1}^{\infty} \frac{a_t^{-1}[H(X_{t-m}^{t-1})]^T\gamma_t H(X_{t-m}^{t-1})Q_t(d_t^T X_{t-m}^{t-1})}{1 + a_{t-1}^{-1}d_t^T\gamma_{t-1}^{-1}d_t} < \infty;$$

then $a_t^{-1}(\hat{\theta}_t - \theta)^T\gamma_t^{-1}(\hat{\theta}_t - \theta)$ converges to a finite limit almost surely.

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

(T3)

$$\sum_{t=1}^{\infty} \inf_{\substack{\epsilon \le |u| \le 1/\epsilon \\ \theta+u \in \bar{U}_{t-1}}} -u^T(a_t^{-1}\gamma_t^{-1} - a_{t-1}^{-1}\gamma_{t-1}^{-1})u - 2a_t^{-1}u^T H(X_{t-m}^{t-1})P_t(u^T X_{t-m}^{t-1}) = \infty$$

on A,

(T4)

$$\sum_{t=1}^{\infty} \sup_{\substack{\epsilon \le |u| \le 1/\epsilon \\ \theta+u \in \bar{U}_{t-1}}} a_t^{-1}[H(X_{t-m}^{t-1})]^T \gamma_t H(X_{t-m}^{t-1})Q_t(u^T X_{t-m}^{t-1}) < \infty \quad \textit{on } A,$$

then $a_t^{-1}(\hat{\theta}_t - \theta)^T\gamma_t^{-1}(\hat{\theta}_t - \theta) \longrightarrow 0$ *(a.s.), for any starting point* $\hat{\theta}_0$.

Proof. Consider Lemma 2.2.1 with $V_t(u) = u^T C_t u$ where $C_t = a_t^{-1}\gamma_t^{-1}$. Since $V_t'(u) = 2u^T C_t$ and $V_t''(u) = 2C_t$, by (T1), we have

$$\begin{aligned}
& [\mathcal{K}_t(u)]^+ \\
= \; & \Big[\Delta V_t(u) + V_t'(u)\gamma_t(\theta+u)R_t(\theta+u) + \eta_t(\theta+u)\Big]^+ \\
= \; & \Big[u^T(a_t^{-1}\gamma_t^{-1} - a_{t-1}^{-1}\gamma_{t-1}^{-1})u + 2u^T a_t^{-1}\gamma_t^{-1}\gamma_t H(X_{t-m}^{t-1})P_t(u^T X_{t-m}^{t-1}) + \eta_t(\theta+u)\Big]^+ \\
\le \; & [\eta_t(\theta+u)]^+,
\end{aligned}$$

where

$$
\begin{aligned}
& \eta_t(\theta+u) \\
= \; & \frac{1}{2}\sup_v E\Big\{\Big[R_t(\theta+u)+\varepsilon_t(\theta+u)\Big]^T\gamma_t(\theta+u)V_t''(v) \\
& \qquad\qquad \gamma_t(\theta+u)\Big[R_t(\theta+u)+\varepsilon_t(\theta+u)\Big]\Big|\mathcal{F}_{t-1}\Big\} \\
= \; & E\Big\{\Big[H(X_{t-m}^{t-1})\varphi_t\Big(X_t-(\theta+u)^TX_{t-m}^{t-1}\Big)\Big]^T\gamma_t a_t^{-1}\gamma_t^{-1} \\
& \qquad\qquad \gamma_t\Big[H(X_{t-m}^{t-1})\varphi_t\Big(X_t-(\theta+u)^TX_{t-m}^{t-1}\Big)\Big]\Big|\mathcal{F}_{t-1}\Big\} \\
= \; & E\left\{a_t^{-1}[H(X_{t-m}^{t-1})]^T\gamma_t H(X_{t-m}^{t-1})\varphi_t^2(\xi_t-u^TX_{t-m}^{t-1})\Big|\mathcal{F}_{t-1}\right\} \\
= \; & a_t^{-1}[H(X_{t-m}^{t-1})]^T\gamma_t H(X_{t-m}^{t-1})E\left\{\varphi_t^2(\xi_t-u^TX_{t-m}^{t-1})\Big|\mathcal{F}_{t-1}\right\} \\
= \; & a_t^{-1}[H(X_{t-m}^{t-1})]^T\gamma_t H(X_{t-m}^{t-1})Q_t(u^TX_{t-m}^{t-1}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\sum_{t=1}^{\infty}\frac{[\mathcal{K}_t(\Delta_{t-1})]^+}{1+V_{t-1}(\Delta_{t-1})} &\le \sum_{t=1}^{\infty}\frac{\eta_t(\theta+\Delta_{t-1})}{1+a_{t-1}^{-1}\Delta_{t-1}^T\gamma_{t-1}^{-1}\Delta_{t-1}} \\
&= \sum_{t=1}^{\infty}\frac{a_t^{-1}[H(X_{t-m}^{t-1})]^T\gamma_t H(X_{t-m}^{t-1})Q_t(\Delta_{t-1}^TX_{t-m}^{t-1})}{1+a_{t-1}^{-1}\Delta_{t-1}^T\gamma_{t-1}^{-1}\Delta_{t-1}} \\
&< \infty.
\end{aligned}
$$

Condition (V2) holds.

Furthermore, since $[a]^- \ge -a$,

$$
\begin{aligned}
& [\mathcal{K}_t(u)]^- \\
\ge \; & -u^T(a_t^{-1}\gamma_t^{-1}-a_{t-1}^{-1}\gamma_{t-1}^{-1})u-2u^Ta_t^{-1}\gamma_t^{-1}\gamma_t H(X_{t-m}^{t-1})P_t(u^TX_{t-m}^{t-1}) \\
& -a_t^{-1}[H(X_{t-m}^{t-1})]^T\gamma_t H(X_{t-m}^{t-1})Q_t(u^TX_{t-m}^{t-1})
\end{aligned}
$$

Then, condition (V3) follows immediately from (T3) and (T4). Now, all the conditions of Lemma 2.2.1 hold and $a_t^{-1}(\hat{\theta}_t - \theta)^T \gamma_t^{-1}(\hat{\theta}_t - \theta) \longrightarrow 0$ (a.s.). ■

Corollary 4.3.2 *Let $\hat{\theta}_t$ be estimators generated by* (4.1.5) *with $U_t = \mathbb{R}^m$ and $H(u) = u$. Suppose that there exists a non-decreasing predictable process $a_t > 0$ such that condition (T2) and (T4) in Corollary 4.3.1 hold and*

(F1) *the pdf $g_t(z)$ is bell-shaped (i.e., unimodal) and symmetric about zero;*

(F2) *for each t, $\varphi_t(v)$ is an odd function such that $\varphi_t(v) > 0$ for $v > 0$ and $\int |\varphi_t(z - v)| g_t(z)dz < \infty$ for all $v \in \mathbb{R}$;*

(F3) *$a_t \gamma_t = M$ eventually, where M is a constant matrix.*

Then $\|\hat{\theta}_t - \theta\|^2$ converges to a finite limit (a.s.).

Furthermore, if the process X_t defined by (4.1.1) *is strongly stationary, then $\|\hat{\theta}_t - \theta\|^2 \longrightarrow 0$ (a.s.), for any admissible truncations U_t and starting point $\hat{\theta}_0$.*

Proof. Consider Corollary 4.3.1 with $U_t = \mathbb{R}^m$ and $H(u) = u$. Condition (V1) holds trivially. It follows from (F1), (F2) and Lemma A.9 that if $v \neq 0$,

$$vP_t(v) = v \int_{-\infty}^{\infty} \varphi_t(z - v)\, g_t(z)dz < 0.$$

Therefore,

$$\begin{aligned} & u^T H(X_{t-m}^{t-1}) E_\theta \left\{ \varphi_t(\xi_t - u^T X_{t-m}^{t-1}) \middle| \mathcal{F}_{t-1} \right\} \\ = \;& u^T X_{t-m}^{t-1} E_\theta \left\{ \varphi_t(\xi_t - u^T X_{t-m}^{t-1}) \middle| \mathcal{F}_{t-1} \right\} \\ = \;& u^T X_{t-m}^{t-1} P_t(u^T X_{t-m}^{t-1}) \\ \leq \;& 0. \end{aligned}$$

By (F3), $u^T(a_t^{-1}\gamma_t^{-1} - a_{t-1}^{-1}\gamma_{t-1}^{-1})u = 0$ eventually. Thus (T1) holds. By the first assertion of Corollary 4.3.1, condition (T1) and (T2) imply that $\|\hat{\theta}_t - \theta\|^2$ converges to a finite limit almost surely.

It is now sufficient to prove that for each $\epsilon \in (0, 1)$,

$$\sum_{t=1}^{\infty} a_t^{-1} \inf_{\substack{\epsilon \le |u| \le 1/\epsilon \\ \theta+u \in U_{t-1}}} -u^T X_{t-m}^{t-1} P_t(u^T X_{t-m}^{t-1}) = \infty, \tag{4.3.2}$$

with the convention that the $\inf_{u \in U} v(u)$ of a function $v(u)$ is 1 whenever $U = \emptyset$. By Lemma A.9, $\inf_{\epsilon \le |u| \le 1/\epsilon} -u^T x P_t(u^T x) > 0$ for any $x \neq 0$. Now, it is easy to see that

$$\inf_{\substack{\epsilon \le |u| \le 1/\epsilon \\ \theta+u \in U_{t-1}}} -u^T x P_t(u^T x) \ge \min\left(\inf_{\epsilon \le |u| \le 1/\epsilon} -u^T x P_t(u^T x),\ 1 \right) > 0$$

for any $x \in \mathbb{R}^{m \times 1} \backslash \{0\}$. Since the process X_t is strongly stationary, it follows from the ergodic theorem that in probability P^θ,

$$\lim_{t\to\infty} \frac{1}{t} I_t \ > 0 \quad \text{and} \quad \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \min\left(\inf_{\epsilon \le |u| \le 1/\epsilon} -u^T X_{s-m}^{s-1} P_s(u^T X_{s-m}^{s-1}),\ 1 \right) \ > 0.$$

These imply that (see e.g., Proposition A4 in [67])

$$\begin{aligned} & M \sum_{t=1}^{\infty} a_t^{-1} \inf_{\substack{\epsilon \le |u| \le 1/\epsilon \\ \theta+u \in U_{t-1}}} -u^T X_{t-m}^{t-1} P_t(u^T X_{t-m}^{t-1}) \\ \ge\ & \sum_{t=1}^{\infty} I_t^{-1} \min\left(\inf_{\epsilon \le |u| \le 1/\epsilon} -u^T X_{t-m}^{t-1} P_t(u^T X_{t-m}^{t-1}),\ 1 \right) \\ =\ & \infty \end{aligned}$$

Thus, (T3) holds. Therefore, all the conditions of Corollary 4.3.1 hold and we have $\|\hat{\theta}_t - \theta\|^2 \longrightarrow 0$ (a.s.). ■

Consider a truncated estimator generated from

$$\hat{\theta}_t = \Phi_{U_t}\left(\hat{\theta}_{t-1} + I_t^{-1} X_{t-m}^{t-1} \varphi_t (X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1})\right), \tag{4.3.3}$$

and

$$U_t = S(\theta_t^{LS}, ct^{\epsilon}) \tag{4.3.4}$$

which is a closed sphere set in $\mathbb{R}^m$ with the center at $\hat{\theta}_t^{LS}$ and radius ct^{ϵ}, where $1/4 < \epsilon < 1/2$, $c > 0$ is a constant and $\hat{\theta}_t^{LS}$ is the recursive least squares estimators defined by (4.2.1) and (4.2.2).

Corollary 4.3.3 *Let $\hat{\theta}_t$ be estimators generated by* (4.3.3) *and* (4.3.4). *Suppose that X_t defined by* (4.1.1) *is strongly stationary and*

(S1) *ξ_t are independent and have finite fourth moments for each t;*

(S2) *there exists $t_0 > 0$ and $\tau > 0$ such that*

$$\sup_{\substack{-\tau \le v \le \tau \\ t > t_0}} Q_t(v) < \infty;$$

(S3) *the function $P_t(v)$ are differentiable at $v = 0$ and*

$$\frac{d}{dv} P_t(v)\Big|_{v=0} \le -\frac{l_{gt}}{2},$$

for large t's;

(S4) *there exists non-decreasing predictable process $a_t > 0$ such that*

$$\sum_{t=1}^{\infty} a_t^{-1} (X_{t-m}^{t-1})^T I_t^{-1} X_{t-m}^{t-1} < \infty.$$

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

Proof. Since X_t is strongly stationary, from the ergodic theorem,

$$\lim_{t\to\infty} \frac{X_t^4}{t} \longrightarrow 0,$$

which implies that $X_{t-m}^{t-1}/t^{1/4} \longrightarrow 0$. It follows from Remark 4.2.3 that $t^\epsilon \|\hat{\theta}_t^{LS} - \theta\| \longrightarrow 0$. So, $\theta \in U_t = S(\hat{\theta}_t^{LS}, ct^{-\epsilon})$ eventually. Then we have $\hat{\theta}_t \in S(\hat{\theta}_t^{LS}, ct^{-\epsilon})$ and we obtain that $t^\epsilon \|\hat{\theta}_t - \theta\| = t^\epsilon \|\Delta_t\| \leq 2c$ eventually, therefore, $t^\epsilon \Delta_t$ is also bounded. Now, since $\epsilon \geq 1/4$, we have

$$\Delta_{t-1}^T X_{t-m}^{t-1} = \Delta_{t-1}^T (t-1)^\epsilon (t-1)^{-\epsilon} X_{t-1} \longrightarrow 0.$$

Now consider Corollary 4.3.1 with $\gamma_t = I_t^{-1}$ and $H(u) = u$. Since X_t is strongly stationary and $C_t = a_t^{-1} I_t$, according to Remark 2.2.3, (V1) holds. By (S2) and (S4), condition (T2) holds immediately. By (S3), we have

$$\begin{aligned}
& u^T(a_t^{-1}\gamma_t^{-1} - a_{t-1}^{-1}\gamma_{t-1}^{-1})u + 2a_t^{-1}u^T H(X_{t-m}^{t-1}) P_t(u^T X_{t-m}^{t-1}) \\
\leq\ & a_t^{-1} u^T (I_t - I_{t-1}) u + 2a_t^{-1} u^T X_{t-m}^{t-1} P_t(u^T X_{t-m}^{t-1}) \\
\leq\ & a_t^{-1} u^T l_{gt} X_{t-m}^{t-1} (X_{t-m}^{t-1})^T u - a_t^{-1} l_{gt} (u^T X_{t-m}^{t-1})^2 \\
=\ & 0.
\end{aligned}$$

Thus, condition (T1) holds. By the first assertion of Corollary 4.3.1, $a_t^{-1}(\hat{\theta}_t - \theta)^T I_t(\hat{\theta}_t - \theta)$ converges to a finite limit almost surely. ■

Remark 4.3.4 *Suppose that ξ_t are i.i.d. and X_t is strongly stationary. Then*

$$\frac{I_t}{t} \longrightarrow G$$

where $G < \infty$ is a positive definite $m \times m$ matrix. According to Lemma A.3, condition (S4) holds if we take $a_t = t^\delta$. It follows from Corollary 4.3.3 that $t^{1-\delta}\|\hat{\theta}_t - \theta\|^2 \longrightarrow 0$ (a.s.) for all $\delta > 0$, if other conditions are satisfied.

Corollary 4.3.5 *Let $\hat{\theta}_t$ be estimators generated by* (4.3.3) *and* (4.3.4)*. Suppose that all conditions in Corollary 4.3.3 are satisfied, ξ_t are i.i.d. r.v.'s and*

(Z1)

$$P_t(v) = P(v) = -l_g v + v^{1+\varepsilon_0} O(1)$$

for some $\varepsilon_0 > 0$ as $v \longrightarrow 0$ where $l_g = l_{gt}$;

(Z2)

$$\int_{-\infty}^{\infty} \Big[\varphi(z-u) - \varphi(z)\Big]^2 g(z)dz \longrightarrow 0$$

as $u \longrightarrow 0$.

Then $\hat{\theta}_t$ is locally asymptotically linear.

Proof. Consider Theorem 2.4.1 with $A_t = \sqrt{t}\mathbf{I}$. Since X_t is strongly stationary, I_t/t converges and condition (E2) holds. Since all conditions of Corollary 4.3.3 are satisfied $\Big($with $a_t = t^\delta\Big)$, $t^{1/2-\delta}(\hat{\theta}_t - \theta) \longrightarrow 0$ for any $\delta > 0$. Then, it is easy to check that condition (E1) holds (see Remark 2.4.5(c) for details). It follows from the proof

of Corollary 4.3.3 that $\Delta_{t-1}^T X_{t-m}^{t-1} \longrightarrow 0$ and $t^{-\frac{1}{4}} X_{t-m}^{t-1} \longrightarrow 0$, we have

$$\begin{aligned} & A_t^{-1}\left[\Delta\gamma_t^{-1}\Delta_{t-1} + \tilde{R}_t(\theta+\Delta_{t-1})\right] \\ = \; & t^{-1/2} X_{t-m}^{t-1}\left[l_g(X_{t-m}^{t-1})^T\Delta_{t-1} + P(\Delta_{t-1}^T X_{t-m}^{t-1})\right] \\ = \; & t^{-1/2} X_{t-m}^{t-1}(\Delta_{t-1}^T X_{t-m}^{t-1})^{1+\varepsilon_0} O(1) \\ = \; & t^{-1-\frac{\delta_0}{2}} X_{t-m}^{t-1}(X_{t-m}^{t-1})^T t^{\frac{1-\delta_0}{2}}\Delta_{t-1}(t^{\frac{1}{2}-\delta_0}\Delta_{t-1}^T t^{\frac{\delta_0}{\varepsilon_0}+\delta_0-\frac{1}{2}} X_{t-m}^{t-1})^{\varepsilon_0} O(1), \end{aligned}$$

for some $0 < \delta_0 < \varepsilon_0/4(1+\varepsilon_0)$.

By the proof of Lemma A.3 in Appendix A,

$$\sum_{t=m+1}^{\infty} \frac{(X_{t-m}^{t-1})^T X_{t-m}^{t-1}}{t^{1+\delta_0/2}} < \infty. \tag{4.3.5}$$

Since $X_{t-m}^{t-1}(X_{t-m}^{t-1})^T$ is positive semi-definite, (4.3.5) leads to

$$\sum_{t=m+1}^{\infty} \frac{X_{t-m}^{t-1}(X_{t-m}^{t-1})^T}{t^{1+\delta_0/2}} < \infty.$$

Also, we have

$$t^{\frac{1}{2}-\delta_0}\Delta_{t-1} = \left(\frac{t}{t-1}\right)^{\frac{1}{2}-\delta_0} (t-1)^{\frac{1}{2}-\delta_0}\Delta_{t-1} \longrightarrow 0$$

and

$$t^{\frac{\delta_0}{\varepsilon_0}+\delta_0-\frac{1}{2}} X_{t-m}^{t-1} \le t^{-\frac{1}{4}} X_{t-m}^{t-1} \longrightarrow 0.$$

Thus,

$$\sum_{t=1}^{\infty} A_t^{-1}\left[\Delta\gamma_t^{-1}\Delta_{t-1} + \tilde{R}_t(\theta+\Delta_{t-1})\right] < \infty.$$

It is now following the Kronecker lemma for matrices that condition (Q1) in Proposition 2.4.2 holds, which leads to condition (E3). $\Big($see Lemma A.6 in Appendix A for the Kronecker lemma with $B_t = A_t^2$ and $\alpha_t = A^{-1}[\Delta\gamma_t^{-1}\Delta_{t-1} + \tilde{R}_t(\hat{\theta}_{t-1})]\Big)$.

On the other hand, since

$$E\{\tilde{\varepsilon}_t(u)|\mathcal{F}_{t-1}\} = E\{\varepsilon_t(u)|\mathcal{F}_{t-1}\} = 0$$

for any $u \in \mathbb{R}^m$, condition (Q2) in Proposition 2.4.3 holds. Also, by (4.3.1) we have for $j = 1, ..., m$,

$$\begin{aligned} & \Big[\tilde{\varepsilon}_t^{(j)}(\theta + \Delta_{t-1}) - \varepsilon_t^{(j)}(\theta)\Big]^2 \\ = \;& X_{t-j}^2\Big[\varphi\Big(\xi_t - \Delta_{t-1}^T X_{t-m}^{t-1}\Big) - P(\Delta_{t-1}^T X_{t-m}^{t-1}) - \varphi(\xi_t)\Big]^2 \\ \leq \;& 2X_{t-j}^2\Big[\varphi\Big(\xi_t - \Delta_{t-1}^T X_{t-m}^{t-1}\Big) - \varphi(\xi_t)\Big]^2 + 2X_{t-j}^2 P^2(\Delta_{t-1}^T X_{t-m}^{t-1}). \end{aligned}$$

Then

$$\begin{aligned} & E\left\{\Big[\tilde{\varepsilon}_t^{(j)}(\theta + \Delta_{t-1}) - \varepsilon_t^{(j)}(\theta)\Big]^2 \Big| \mathcal{F}_{t-1}\right\} \\ \leq \;& 2X_{t-j}^2 \int_{-\infty}^{\infty} \Big[\varphi\Big(z - \Delta_{t-1}^T X_{t-m}^{t-1}\Big) - \varphi(z)\Big]^2 g(z)dz + 2X_{t-j}^2 P^2(\Delta_{t-1}^T X_{t-m}^{t-1}). \end{aligned}$$

Since $\Delta_{t-1}^T X_{t-m}^{t-1} \longrightarrow 0$ and $0 < t^{-1}\sum_{s=1}^t X_{s-j}^2 < \infty$, by (Z2) and the Toeplitz lemma, we have

$$t^{-1}\sum_{s=1}^{t} X_{s-j}^2 \int_{-\infty}^{\infty} \Big[\varphi\Big(z - \Delta_{s-1}^T X_{s-m}^{s-1}\Big) - \varphi(z)\Big]^2 g(z)dz \longrightarrow 0$$

and

$$t^{-1}\sum_{s=1}^{t} X_{s-j}^2 P^2(\Delta_{s-1}^T X_{s-m}^{s-1}) \longrightarrow 0.$$

So, (Q3) holds. According to Proposition 2.4.3, condition (E4) in Theorem 2.4.1 holds. Therefore, all the conditions of Theorem 2.4.1 hold which imply that $\hat{\theta}_t$ is locally asymptotically linear. ■

Example 4.3.6 Suppose that ξ_t are independent Student random variables with α degrees of freedom. Consider the following recursive likelihood procedure

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} - I_t^{-1} X_{t-m}^{t-1}\frac{g'}{g}(X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1})\Big) \tag{4.3.6}$$

The pdf of ξ_t is

$$g(z) = C_\alpha \left(1 + \frac{z^2}{\alpha}\right)^{\frac{-\alpha-1}{2}}$$

where $C_\alpha = \mathbf{\Gamma}((\alpha+1)/2)/(\sqrt{\pi\alpha}\ \mathbf{\Gamma}(\alpha/2))$. Since

$$\frac{g'(z)}{g(z)} = -(\alpha+1)\frac{z}{\alpha+z^2},$$

we have

$$\begin{aligned} l_{gt} = l_g &= \int \left(\frac{g'(z)}{g(z)}\right)^2 g(z)\,dz \\ &= C_\alpha(\alpha+1)^2 \int \frac{z^2\,dz}{(\alpha+z^2)^2(1+\frac{z^2}{\alpha})^{\frac{\alpha+1}{2}}} \\ &= C_\alpha \frac{(\alpha+1)^2}{\sqrt{\alpha}} \int \frac{z^2\,dz}{(1+z^2)^{\frac{\alpha+5}{2}}} \\ &= C_\alpha \frac{(\alpha+1)^2}{\sqrt{\alpha}} \frac{\sqrt{\pi}\mathbf{\Gamma}((\alpha+5)/2-3/2)}{2\mathbf{\Gamma}((\alpha+5)/2)} \\ &= \frac{\alpha+1}{\alpha+3}. \end{aligned}$$

We may rewrite (4.3.6) as

$$\hat{\theta}_t = \Phi_{U_t}\Big(\hat{\theta}_{t-1} + \hat{I}_t^{-1}(\alpha+3)X_{t-m}^{t-1}\frac{X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1}}{\alpha + (X_t - \hat{\theta}_{t-1}^T X_{t-m}^{t-1})^2}\Big) \tag{4.3.7}$$

where $\hat{I}_t = \sum_{s=1}^{t} X_{s-m}^{s-1}(X_{s-m}^{s-1})^T + c\mathbf{I}$ or in a recursive form as (4.2.2).

Suppose that $\alpha \geq 5$, $\{X_t\}$ is strongly stationary and (4.3.7) is truncated by $U_t = S(\hat{\theta}_t^{LS}, t^\epsilon)$, where $1/4 < \epsilon < 1/2$. Then $t^{1/2-\delta}\|\hat{\theta}_t - \theta\|$ converges to a finite limit almost surely for all $\delta > 0$. Indeed, consider Corollary 4.3.3 with $\varphi_t = g'/g$. The random innovation ξ_t has a finite fourth moment since $\alpha \geq 5$, $\lim_{v\to 0} P_t(v)/v = -l_{gt} \leq -l_{gt}/2$, and (S4) holds with $a_t = t^\delta$ for any $\delta > 0$. Furthermore, conditions of Corollary 4.3.5 are also satisfied, and $\hat{\theta}_t$ is asymptotic linear.

4.4 Summary

Recursive estimation for AR processes has been studied by a number of authors (see e.g., Lai and Ying [45] and Chen [15] and [16]). However, these methods where mostly focused on linear cases, resulting in recursive least squares type procedures. The class of recursive parameter estimators considered in this chapter covers non-linear cases, and also makes it possible to incorporate auxiliary information into the estimation process, by considering truncations with moving bounds. This class was introduced and studied by Sharia [68] for AR(1) processes. In this chapter, we applied the results obtained in the previous chapters to study similar procedures for AR(m) processes. In particular, convergence, rate of convergence and asymptotic linearity of recursive estimators are established for multi-dimensional AR models .

Two important cases are considered in detail: the recursive least squares (RLS), and the recursive likelihood with the RLS truncations.

The results for the RLS are derived from the corollaries presented in Section 3.2.3. This section demonstrates that, the conditions are minimal in the sense that they do not impose any additional restrictions when applied to the well known models. Although the asymptotic behaviour of the RLS is well known, to our best knowledge, convergence and the rate of convergence of the RLS obtained in this chapter cannot be derived from any other result in Stochastic Approximation. For example, Lai and Ying [45] study procedures similar to the RLS above for general linear time series models. However, the result given by [45] for AR models is more restrictive than the one in this thesis.

In Section 4.3, the RSL are used as an auxiliary estimator to carry out the truncations. This guarantees the convergence of the truncated likelihood type procedure, and given an appropriate choice of the step-size sequence, leads to the asymptotically efficient estimation.

Chapter 5

Simulations

Monte-Carlo simulations for the RM type SA procedures are presented in the following three specific cases: polynomials with integer degrees, estimation of the shape parameter of the Gamma distribution, and estimation in Autoregressive model of order 2.

5.1 Finding roots of polynomials

Let us consider a problem described in Example 2.5.13 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.

Figure 5.1 shows 30 steps of 3 estimators generated from procedure (2.5.1) with $a_t = 3t$ and starting points at -2, 0 and 5 respectively, where the root $z^0 = 2$ and truncation sequence $U_t = [-\log 3t, \log 3t]$. As we can see, the estimators go

towards the root following a zigzag path. 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, and Figure 5.1 gives us an illustration of how it works.

Also, using Corollary 2.5.11 (with $\frac{1}{3}R(z)$ instead of $R(z)$), we obtain that our estimators are asymptotically linear and since the error terms are i.i.d., it follows that the estimators are asymptotically normal. A histogram of estimators at iterate 30 over 500 replications (with $Z_0 = 0$) is shown in Figure 5.2.

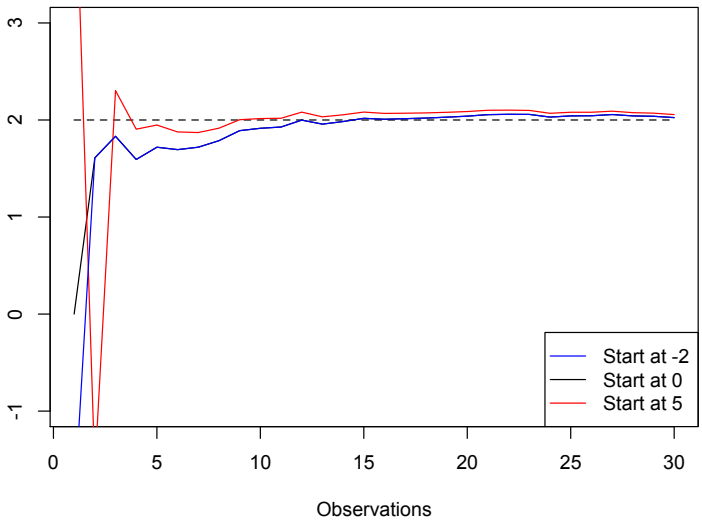

Figure 5.1: Realizations of the estimator (2.5.1)

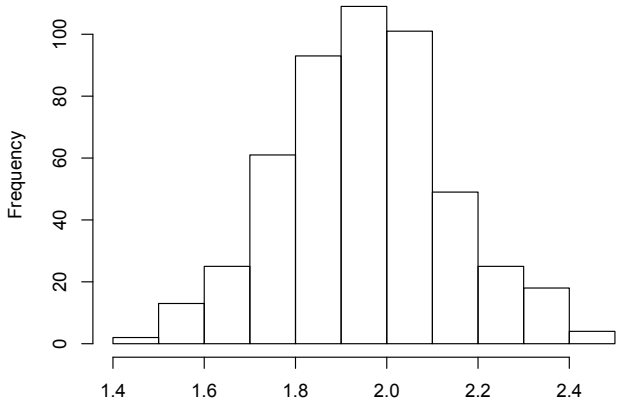

Figure 5.2: Histogram of estimator at step 30

5.2 Estimation of the shape parameter of the Gamma distribution

Let us consider procedure (3.2.2) in Example 3.2.1 with following two sets of truncations $U_t = [\alpha_t, \beta_t]$

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

(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 5.3 shows realizations of procedures (3.2.2) 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, this might be due to singularity at 0 of the functions appearing in the procedure. Increasing α might

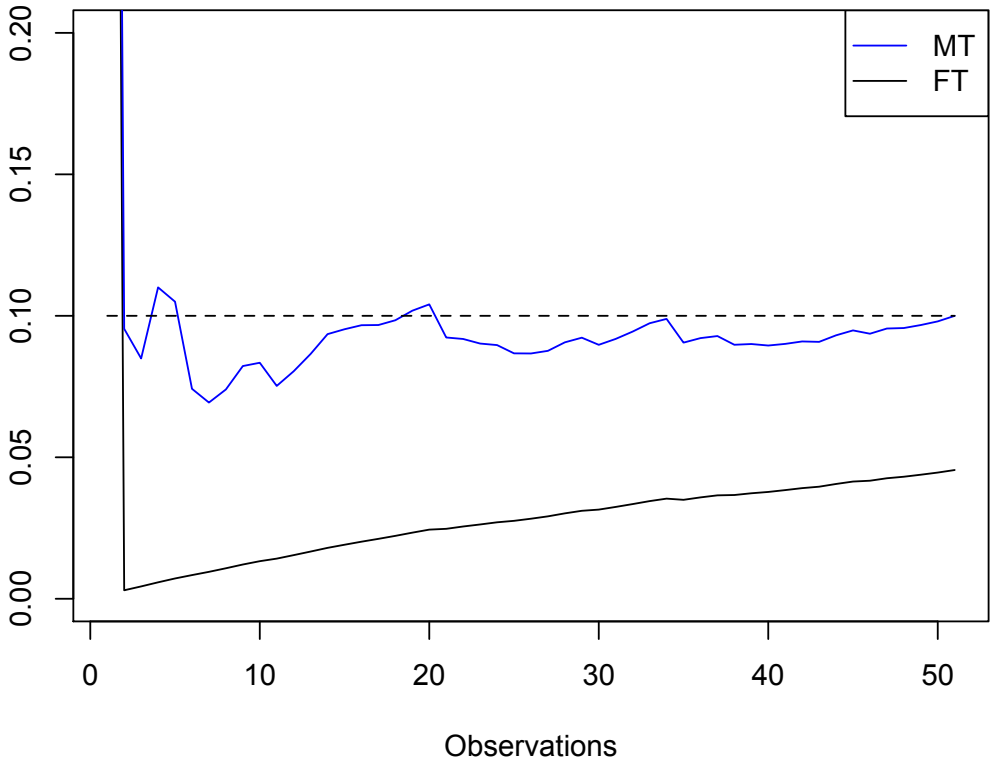

Figure 5.3: Performance of procedure (3.2.2)

fix the problem. However, unless we have additional information about location of the parameter θ, we need to use the recursive procedures with moving truncations.

5.3 An AR(2) example

Let us consider the process

$$X_t = -0.9X_{t-1} - 0.5X_{t-2} + \xi_t.$$

where ξ_t are i.i.d. Student random variables with 5 degrees of freedom.

Let us consider the following recursive estimators:

(1) RLS – the recursive least squares estimator defined by procedure (4.2.1) in Section 4.2.

(2) RML – the recursive maximum likelihood estimator defined by procedure (4.1.4).

(3) RMLtri – the RML estimator truncated in the stationarity region (4.1.6) of AR(2) which is a triangular region on $\mathbb{R}^2$.

(4) RMLls – the RML estimator truncated by RLS, which is defined by procedure (4.3.3) and (4.3.4). The estimator is forced to the nearest point in the sphere whose center is the current RLS and radius is ct^{ε}, where c is a positive tuning constant.

All the estimators are simulated with starting point $(0.4, -0.6)$, $c = 1.5$, $\varepsilon = -1/3$ and $\hat{I}_0 = \mathbf{I}$.

Figure 5.4 shows realizations of the estimating procedures for 50 steps. As we can see, all the estimators go up and down around the true parameters. Simulations show that typical realizations of all 4 graphs are very similar.

Figures 5.5 shows the MSEs for the first 50 steps (observations) and Figure 5.6 shows the MSEs of the estimators form the step 400 to 500 based on 50 replications. As we can see, RMLls has the smallest MSE. Therefore, it is more efficient than RLS. This demonstrates the idea of moving bound truncations, that is, one can force a convergent but not necessarily efficient estimator to the true parameter by truncations, and this truncated estimator may have both convergence and efficiency.

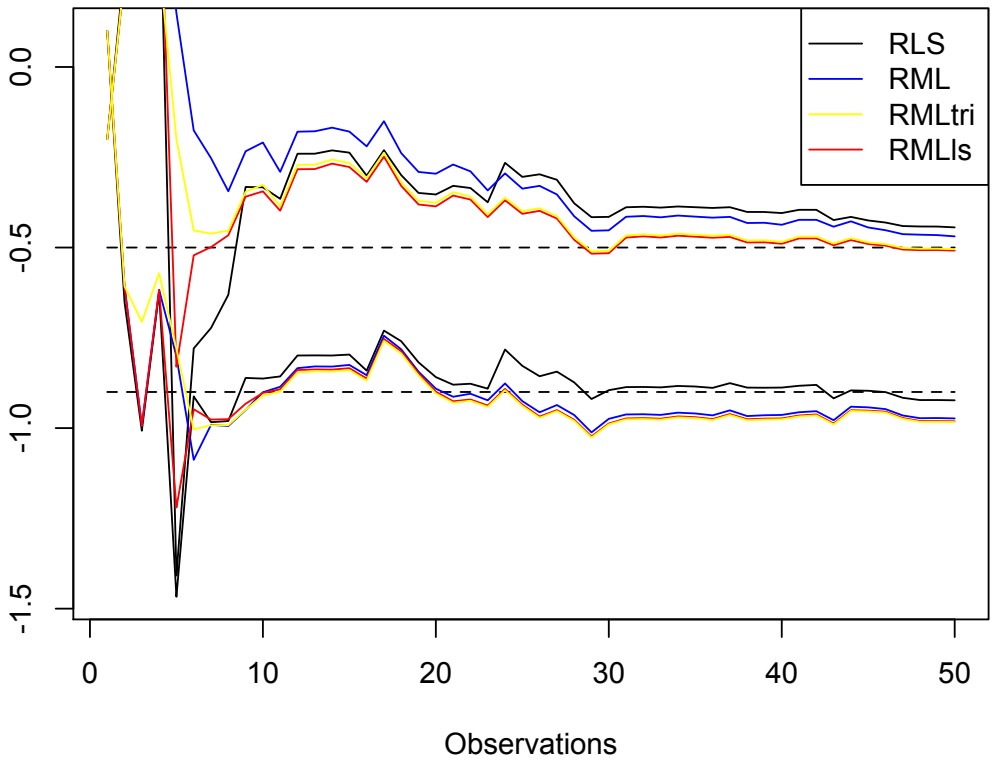

Figure 5.4: Realizations of the estimator

Remark 5.3.1 Step-size sequences suggested in Remark 2.4.4 and in Section 3.1 have been derived from the asymptotic considerations. In practice, especially if the number of observations is small or moderately large, behaviour of step-size sequences for the first several steps might also be important. According to Remark 2.4.4, to achieve asymptotic linearity, we have to choose a step-size sequence in such a way that $\triangle\gamma_t(z) \approx -R'_t(z)$ for large t's. So, we can consider any sequence of the form $C + c_t\gamma_t$, where c_t is non-negative with $c_t = 1$ for large t's, and C is a constant. In practice, we can treat c_t and C as tuning constants to manage behaviour of

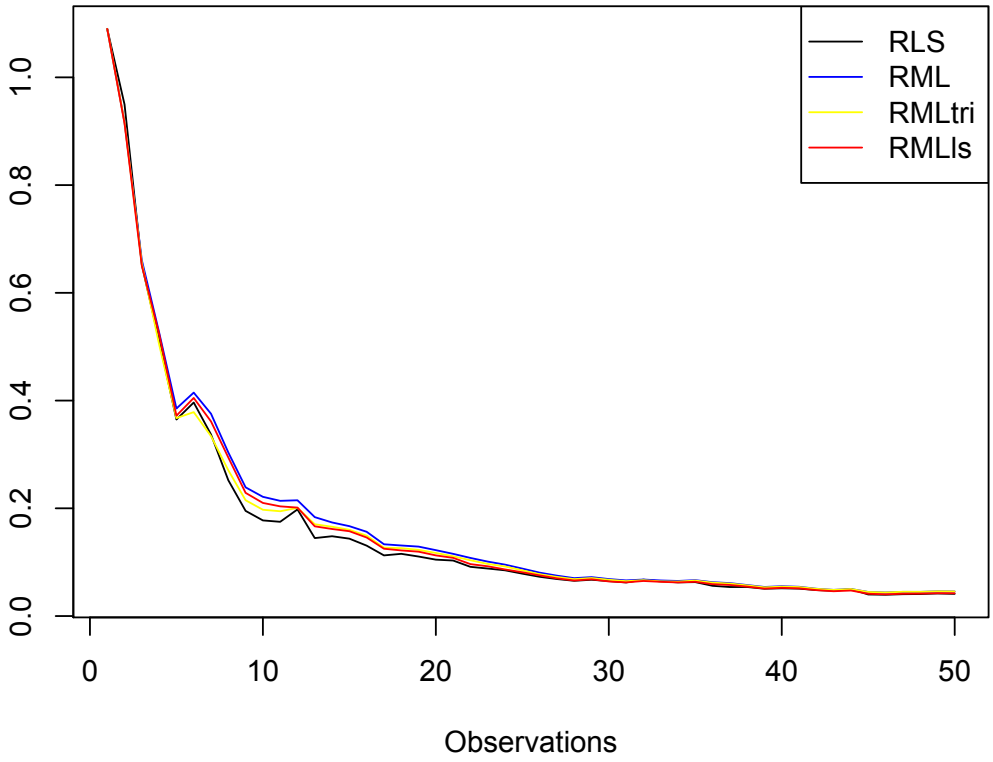

Figure 5.5: Mean Square Errors of each process

the procedure for the first several steps (see Sharia [67], Remark 4.4). A reasonable shape of the graph against t is similar to those in Figure 5.4, that is, some oscillation at the beginning of the procedure and then settling down at a particular level.

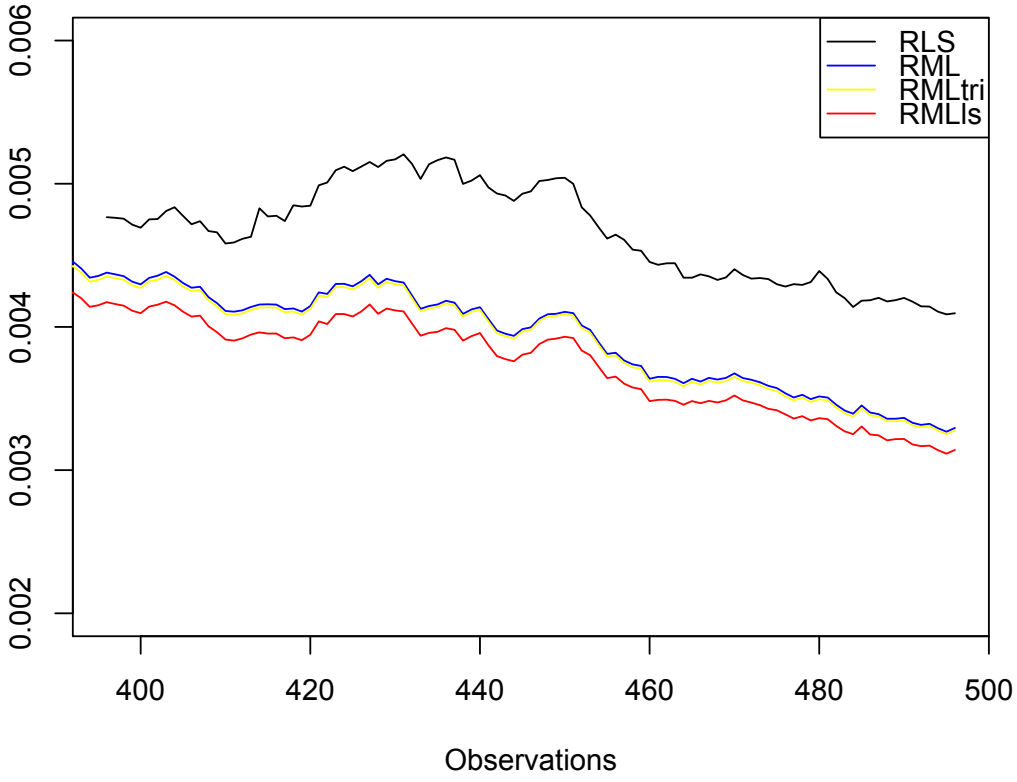

Figure 5.6: Mean Square Errors of each process

Chapter 6

Conclusions

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

Three main asymptotic properties of the RM type SA are established: convergence, rate of convergence, and asymptotic linearity.

Convergence and rate of convergence of the RM type SA are studied using the

Robbins-Siegmund Lemma and by considering time dependent random Lyapunov type functions.

It is also shown, that under quite mild conditions, the SA process 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. Furthermore, these results help to identify step-size sequences that are optimal for a given set of R_t functions.

The above results have been applied to deterministic and time homogeneous regression functions. The results demonstrate that conditions we used in the thesis are minimal in the sense that they do not impose any additional restrictions when applied to the classical case. Furthermore, Section 2.5 contains new results even for the classical SA problem. In particular, truncations with moving bounds give a possibility to use SA in the cases when classical conditions on the function R do not hold. Also, a very 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. This observation would not surprise experts working in this field. However, we failed to find it in a written form in the existing literature.

Applications of the theoretical results of the thesis to the problems of parametric statistical estimation for various statistical models are also presented. A particular attention is given to the on-line estimation of the parameters for AR(m) processes. Two important cases are considered in detail: the recursive least squares (RLS), and the recursive likelihood with the RLS truncations.

Finally, Monte-Carlo simulations are also presented for some specific cases.

The material has been arranged in five chapters. Each chapter contains a brief

introduction and a summary to explain novelty of the results presented in a given chapter. Main lemmas and theorems are followed by various corollaries and remarks containing sufficient conditions for the convergence and explaining some of the assumptions. These corollaries are presented in such a way, that each subsequent corollary imposes conditions that are more restrictive than the previous one.

We believe that the results of the thesis are of a publishable quality. One paper, based on the material presented in Chapters 2 and 3, is almost ready for publication. The second paper will contain results presented in Chapter 4. In the nearest future, we also plan to study parameter estimation for the exponential family of Markov chains presented in Chapter 3 in more details.

Appendix A

Lemmas and Propositions

Lemma A.1 *Let $\mathcal{F}_0$, $\mathcal{F}_1$, ... be an non decreasing sequence of σ-algebras and X_n, β_n, ξ_n, $\zeta_n \in \mathcal{F}_n$, $n \geq 0$, be nonnegative random valuables such that*

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

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 [60]. Note also that this lemma is a special case of the theorem on the convergence sets of nonnegative semi-martingales (see, e.g., Lazrieva et al [46]). ■

Lemma A.2 *Let* $\{\alpha_t\}$ *be a sequence of real* $m \times 1$ *column vector and* $I_t = \mathbf{I} + \sum_{s=1}^{t} \alpha_s \alpha_s^T$. *Then*

$$\alpha_t^T I_t^{-1} \alpha_t \leq 1.$$

Proof. *Denote* $\beta_t = I_t^{-1}\alpha_t$, *then* $\alpha_t = I_t\beta_t$ *and*

$$\alpha_t^T I_t^{-1}\alpha_t = (I_t\beta_t)^T\beta_t = \beta_t^T I_t\beta_t \geq \beta_t^T\alpha_t\alpha_t^T\beta_t = (\alpha_t^T\beta_t)^2 = (\beta_t^T I_t\beta_t)^2 = (\alpha_t^T I_t^{-1}\alpha_t)^2.$$

So,

$$\alpha_t^T I_t^{-1}\alpha_t \geq (\alpha_t^T I_t^{-1}\alpha_t)^2$$

and this implies that $\alpha_t^T I_t^{-1}\alpha_t \leq 1$. ■

Lemma A.3 *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 real process* κ_t *satisfying:*

$$\frac{I_t}{\kappa_t} \to G,$$

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. $tr(I_t) = m + \sum_{s=1}^{t} \alpha_s^T\alpha_s$ is a non-decreasing sequence of positive numbers, we have (see Proposition A2 in Sharia [66])

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

Since $\frac{I_t}{\kappa_t}$ convergent, $\frac{tr(I_t)}{\kappa_t}$ 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,

$$\kappa_t I_t^{-1} \to G^{-1} \implies \kappa_t \lambda_t^{max} \quad \text{convergent,}$$

where λ_t^{max} is the largest eigenvalue of I_t^{-1}, then

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

■

Lemma A.4 *(The Toeplitz Lemma)*

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

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

Proof. Proof can be found in Loève [52] (P.250). ■

Lemma A.5 *Let* A *and* B *be* $m \times m$ *invertible matrices and there exists vector* α *such that* $A = B + \alpha\alpha^T$*. Then*

$$A^{-1} = B^{-1} - B^{-1}\alpha(1 + \alpha^T B^{-1}\alpha)\alpha^T B^{-1}$$

Proof. Since

$$\begin{aligned}
& (1-\alpha^T A^{-1}\alpha)(1+\alpha^T B^{-1}\alpha) \\
= \; & 1-\alpha^T A^{-1}\alpha+\alpha^T B^{-1}\alpha-\alpha^T A^{-1}\alpha\alpha^T B^{-1}\alpha \\
= \; & 1-\alpha^T A^{-1}\alpha+\alpha^T B^{-1}\alpha-\alpha^T A^{-1}(A-B)B^{-1}\alpha \\
= \; & 1-\alpha^T A^{-1}\alpha+\alpha^T B^{-1}\alpha+\alpha^T A^{-1}\alpha-\alpha^T B^{-1}\alpha \\
= \; & 1,
\end{aligned}$$

we have

$$\begin{aligned}
\alpha\alpha^T &= \alpha\alpha^T(1-\alpha^T A^{-1}\alpha)(1+\alpha^T B^{-1}\alpha) \\
&= (\alpha\alpha^T-\alpha\alpha^T A^{-1}\alpha\alpha^T)(1+\alpha^T B^{-1}\alpha) \\
&= (B-A+2\alpha\alpha^T-\alpha\alpha^T A^{-1}\alpha\alpha^T)(1+\alpha^T B^{-1}\alpha) \\
&= [B-(A-\alpha\alpha^T)A^{-1}(A-\alpha\alpha^T)](1+\alpha^T B^{-1}\alpha) \\
&= (B-BA^{-1}B)(1+\alpha^T B^{-1}\alpha).
\end{aligned}$$

Then

$$B^{-1}\alpha\alpha^T B^{-1} = (B^{-1}-A^{-1})(1+\alpha^T B^{-1}\alpha)$$

and

$$B^{-1}-A^{-1} = B^{-1}\alpha(1+\alpha^T B^{-1}\alpha)^{-1}\alpha^T B^{-1}.$$

■

Lemma A.6 *(The Kronecker Lemma for Matrices)*

Let α_t be a sequence of real $m\times 1$ vectors for which $\|\sum_{t=1}^{\infty}\alpha_t\|<\infty$, and B_t be a sequence of $m\times m$ positive definite real matrices such that B_t-B_{t-1} is non-negative

definite for all $t \in \mathbb{Z}^+$, $tr^{-1}(B_t) \longrightarrow 0$ *as* $t \longrightarrow 0$ *and* $\lambda^{max}(B_t)/\lambda^{min}(B_t)$ *is bounded. Then*

$$\lim_{t\to\infty} B_t^{-1} \sum_{s=1}^{t} B_s \alpha_s = 0.$$

Proof. A proof can be found in Anderson and Moore [1].

Proposition A.7 *Consider a closed sphere* $U = S(\alpha, r)$ *in* $\mathbb{R}^m$ *with center at* $\alpha \in \mathbb{R}^m$ *and 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*

$$\lambda_C^{max} v^2 \leq \lambda_C^{min} r^2, \tag{A.1}$$

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

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

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

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

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

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

and hence

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

Therefore

$$
\begin{aligned}
(z'-z_0, z-z')_C &= (z'-\alpha, z-z')_C + (\alpha - z_0, z-z')_C \\
&\frac{1-\delta}{\delta}\,\|z'-\alpha\|_C^2 - \frac{1-\delta}{\delta}\,(z_0-\alpha, z'-\alpha)_C \ge 0 \qquad \text{(A.3)}
\end{aligned}
$$

due to (A.2). Since

$$
z' - z_0 = z - z_0 - (z - z'),
$$

we get

$$\|z' - z_0\|_C^2 = \|z - z_0\|_C^2 + \|z - z'\|_C^2 - 2(z - z_0, z - z')_C$$
$$= \|z - z_0\|_C^2 + \|z - z'\|_C^2 - 2\|z - z'\|_C^2 - 2(z' - z_0, z - z')_C$$
$$= \|z - z_0\|_C^2 - \|z - z'\|_C^2 - 2(z' - z_0, z - z')_C \leq \|z - z_0\|_C^2$$

due to (A.3). ■

Proposition A.8 *If d_t is a nondecreasing sequence of positive numbers such that $d_t \longrightarrow +\infty$, then*

$$(a) \quad \sum_{t=1}^{\infty} \Delta d_t / d_t = +\infty$$

and

$$(b) \quad \sum_{t=1}^{\infty} \Delta d_t / d_t^{1+\varepsilon} < +\infty$$

for any $\varepsilon > 0$.

Proof. These can easily be obtained by elementary arguments (see, e.g., Sharia [66], Appendix 2). ■

Lemma A.9 *Suppose that $g \not\equiv 0$ is a nonnegative even function on $\mathbb{R}$ and $g \downarrow 0$ on $\mathbb{R}_+$. Suppose also that φ is a measurable odd function on $\mathbb{R}$ such that $\varphi(z) > 0$ for $z > 0$ and $\int_{\mathbb{R}} |\varphi(z - w)| g(z) dz < \infty$ for all $w \in \mathbb{R}$. Then*

$$w \int_{-\infty}^{\infty} \varphi(z - w)\, g(z) dz < 0$$

for any $w \neq 0$. Furthermore, if $g(z)$ is continuous, then for any $\varepsilon \in (0, 1)$

$$\sup_{\varepsilon \leq |w| \leq 1/\varepsilon} w \int_{-\infty}^{\infty} \varphi(z - w)\, g(z) dz < 0.$$

Proof. The proof of this lemma is given in Sharia [67] (Lemma A.2 in Appendix A). ■

Lemma A.10 *Let $b \geq a > 0$. Then*

$$\frac{b-a}{a} \geq \ln b - \ln a.$$

proof. We have

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

Proposition A.11 *Suppose a_t, $t \in \mathbb{N}$ is a nondecreasing 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

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

The latter follows easily from Lemma A.10:

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

It is clear that Proposition A.11 applies, e.g., to $a_t = t^{\varepsilon}$ with any $\varepsilon > 1$.

Appendix B

Properties of Gamma distribution

This appendix is from Sharia [70]. In Example 3.2.1, we will need the following properties of the Gamma function (see, e.g., Whittaker and Watson [76], 12.16). $\log'\mathbf{\Gamma}$ is increasing, $\log''\mathbf{\Gamma}$ is decreasing and continuous, and

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

The latter implies that

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

and

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

Also (see Cramer [18], 12.5.4),

$$\log'\mathbf{\Gamma}(x) \leq \ln(x). \tag{B.3}$$

Then,

$$E_\theta\left\{\log X_1\right\} = \log'\mathbf{\Gamma}(\theta) \quad \text{and} \quad E_\theta\left\{(\log X_1)^2\right\} = \log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta))^2 \tag{B.4}$$

and

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

Let us show that the conditions of Corollary 2.3.4 hold. Denote $\psi_t(u) = R(u)+\varepsilon_t(u)$, since

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

using (B.4) and (B.2) we obtain

$$\frac{E\left\{\|\psi_t(u)\|^2 \mid \mathcal{F}_{t-1}\right\}}{1+\|u-\theta\|^2} = \frac{\log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u))^2}{(\log''\mathbf{\Gamma}(u))^2(1+\|u-\theta\|^2)} \tag{B.5}$$

$$\leq \frac{u^2}{1+(u-\theta)^2}\left(\log''\mathbf{\Gamma}(\theta) + (\log'\mathbf{\Gamma}(\theta) - \log'\mathbf{\Gamma}(u))^2\right).$$

Now, $u^2/(1+(u-\theta)^2) \leq C$. Here and further on in this subsection, C denotes various constants which may depend on θ. So, using (B.3) we obtain

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

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

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

Condition (H2) and (H3) of Corollary 2.5.1 is now immediate from the second part of (3.2.4). It remains to check that (D3) of Corollary 2.3.5 holds. Indeed,

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

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

$$\inf_{\substack{\varepsilon \le \|u-\theta\| \le 1/\varepsilon \\ u \in U_{t-1}}} -(u-\theta)R(u) \ge \frac{\inf_{\varepsilon \le \|u-\theta\| \le 1/\varepsilon} \left(\log' \mathbf{\Gamma}(u) - \log' \mathbf{\Gamma}(\theta)\right)(u-\theta)}{\sup_{u \in U_{t-1}} \log'' \mathbf{\Gamma}(u)} \ge \frac{C}{\log'' \mathbf{\Gamma}(\alpha_{t-1})} \tag{B.6}$$

where C is a constant that may depend on ε and θ. Since $\alpha_{t-1} < 1$ for large t's, it follows (B.1) that $1/\log'' \mathbf{\Gamma}(\alpha_{t-1}) \ge \alpha_{t-1}^2/2$. Condition (D3) of Corollary 2.3.5 is now immediate from the first part of (3.2.4).

Note that with $\beta_t = \infty$ the procedure fails condition (2) of Corollary 2.3.4. Indeed, (B.5) and (B.1) implies that

$$\sup_{\alpha_t \le u} \frac{E\left\{\psi_t^2(u) \mid \mathcal{F}_{t-1}\right\}}{1+(u-\theta)^2} \ge \sup_{\alpha_t \le u} \frac{\left\{\log'' \mathbf{\Gamma}(\theta) + \left(\log' \mathbf{\Gamma}(\theta) - \log' \mathbf{\Gamma}(u)\right)^2\right\} u^4}{(1+u)^2(1+(u-\theta)^2)} = \infty \tag{B.7}$$

Appendix C

Codes of Monte-Carlo Simulations

Codes of Section 5.1

```
l=7                    ## Highest order of the polynomial
p=0                    ## Start point
df=7                   ## Degrees of freedom of the innovation
n=30 -1                ## Iterates
z=2                    ## True parameter
R=500                  ## No. of replications
Final=0                ## Define the set for Final estimators
Plots=1                ## 1--plot estimators, 2--hist

for(j in 1:R){         ## Outer loop
theta=p
po=p
for (i in 1:n){        ## Inner loop
xi=rt(1,df)            ## Random innovation
po=po+(i)^(-1)*(-(po-z)-(po-z)^l+2*(po-z)^(l-1)-5*(po-z)^(l-2)+xi)

if (po>log(i)){po=log(i)}      ## moving truncations which is indepent to the order
if (po< -log(i)){po=-log(i)}

theta=c(theta,po)
}                              ##End inner loop
Final=c(Final,po)
}                              ##End outer loop
Final=Final[2:(R+1)]

if(Plots==2){
hist(Final,main="",xlab="")
}
if(Plots==1){
plot(theta,,'l' ,
xlab="Observations",ylab="")
lines(rep(z,(n+1)),lty=2)
}
###
```

Codes of Section 5.2

```
s= 1                ## Start point
K=0.1               ## True parameter
n=50                ## Iterates
a=0.003             ## Fixed truncation LOWER bound
b=100               ## Fixed truncation UPPER bound
R=100               ## No. of replications
Plots=1             ## 1--plot estimators, 2--hist
FinalF=0            ## Define the set of final Fix-Bounded estimators
FinalM=0            ## Define the set of final Moving-Bounded estimators
C=0.1               ## Moving bound constant
T=0                 ## Sum of Record of the last iterate when truncation works
CI=1                ## Constant before the normalizing process

for(j in 1:R){      ## Outer loop
thetaF=s
thetaM=s
gF=s
gM=s
for (i in 1:n){ ## Inner loop

xi=rgamma(1,K,1)
gF=gF+CI*(i*trigamma(gF))^(-1)*(log(xi)-digamma(gF))

if(gF<a){gF=a}
if(gF>b){gF=b}
#xi=rgamma(1,K,1)
gM=gM+CI*(i*trigamma(gM))^(-1)*(log(xi)-digamma(gM))

if(gM<C*(log(i+2))^(-1/2)){gM=C*(log(i+2))^(-1/2);t=i}
if(gM>(i+2)){gM=(i+2);t=i}

thetaF=c(thetaF,gF)
thetaM=c(thetaM,gM)
}                           ## End inner loop
T=T+ t
FinalF=c(FinalF,gF)
FinalM=c(FinalF,gM)
}                           ## End outer loop

par(mfrow=c(1,1))
FinalF=FinalF[2:(R+1)]
FinalM=FinalM[2:(R+1)]
if(Plots==2){
par(mfrow=c(1,2))
hist(FinalF,main="",xlab="")
hist(FinalM,main="",xlab="")
}
if(Plots==1){
plot(thetaF,,'l' , xlab="Observations",ylab="",ylim=c(0,0.2))
lines(thetaM,col="blue")
lines(rep(K,(n+1)),lty=2)
legend("topright", col = c("blue", "black"),
legend = c("MT", "FT"),
lty = 1, merge = TRUE)
}
```

T/R

Codes of Section 5.3

```
t1=-0.9
t2=-0.5
theta=c( t1, t2 )                  ## True value of the parameter
s1=0.1                             ## Start value of theta(1)
s2=-0.2                            ## Start value of theta(2)
n=500 +1                           ## No. of Iterates
R=50                               ## No. of replications
df=5                               ## Degree of Freedom
CI=1                               ## Initial enties of I_0, the mornalization process
q=10 #%                            ## First q% interats take constant K times the mornalization process
K=1
e=-1/3 ## epsilon, in charge of the shrinking rate of truncation process
C=1.5 ## Constant in truncations
t=0 ## t will be the latest iterates when the truncation works
T=0 ## Sum of t over n replication
Plots=4 ## 1--plot the MEAN, 2--plot the MSE, 3--both, 4--MSE for the last 100 iterates

######################### Define arraies of Mean Square Errors
MSEmlt=matrix(c(rep(0,(n-1))),1,(n-1))      ## Maximum Likelihood Truncated by trianglur region
MSEmle=matrix(c(rep(0,(n-1))),1,(n-1))      ## Maximum Likelihood
MSElse=matrix(c(rep(0,(n-1))),1,(n-1))      ## Least squares
MSEmltls=matrix(c(rep(0,(n-1))),1,(n-1))    ## Maximum Likelihood Truncated by LS

######################### Define arraies of Means
MEANmlt=matrix(c(rep(0,(n*2-2))),2,(n-1))   ## Maximum Likelihood Truncated by trianglur region
MEANmle=matrix(c(rep(0,(n*2-2))),2,(n-1))   ## Maximum Likelihood
MEANlse=matrix(c(rep(0,(n*2-2))),2,(n-1))   ## Least squares
MEANmltls=matrix(c(rep(0,(n*2-2))),2,(n-1)) ## Maximum Likelihood Truncated by LS

for (j in 1:R){ ################################### Start outer loop
######################### AR process
u=c(rep(0,n+101))
for (i in 3:(n+101)){
u[i]=theta[1]*u[i-1]+theta[2]*u[i-2]+rt(1,df)
}
u=u[100:(n+101)]

######################### initial variables of \hat I_0^{-1}
im=matrix(c(1,0,0,1),2,2) * CI

######################### Define arraies of estimators
MLT=as.matrix(c(s1,s2))          ## Maximum Likelihood Truncated by trianglur region
MLE=as.matrix(c(s1,s2))          ## Maximum Likelihood
LSE=as.matrix(c(s1,s2))          ## Least squares
MLTLS=as.matrix(c(s1,s2))        ## Maximum Likelihood Truncated by LS

################## Start points
tml=MLT
ls=LSE
ml=MLE
tml2=MLTLS

################## ################## Start inner loop
for (i in 3:n){

################## Update \hat I_i^{-1}
```

```
im=im-im%*%as.matrix(u[(i-1):(i-2)]) %*% ( 1/(1+ u[(i-1):(i-2)]%*%im%*%as.matrix(u[(i-1):(i-2)])))%*%u[(i-1):(i-2)] %*%im

if(i<(q*n)){im1=im*K} else{im1=im} ##### In order to apply q and K.

################## Triangle Truncated MLE
tml=tml+(df+3)*im1%*% as.matrix(u[(i-1):(i-2)]) %*%((u[i]-u[(i-1):(i-2)]%*%tml)/(df+(u[i]-u[(i-1):(i-2)]%*%tml)^2))

x=tml[1]
y=tml[2]

if (y<=x+1&y>=x-3&y>=1-x){
y1=y-(y+x-1)/2
x1=x-(x+y-1)/2
}
else{
if (y>=x+1&y>=-x-3&y<=1-x){
y1=y-(y-x-1)/2
x1=x+(y-x-1)/2
}else{

if (y<=-1&x>=-2&x<=2){
y1=-1
x1=x
}else{

if (y>=x+1&y>=1-x){
y1=1
x1=0
}else{

if (x>=2&y<=x-3){
y1=-1
x1=2
}else{

if (x<=-2&y<=-3-x){
y1=-1
x1=-2
}
else {
y1=y
x1=x
}}}}}}
tml[1]=x1
tml[2]=y1

################# MLE
ml=ml+(df+3)*im1%*% as.matrix(u[(i-1):(i-2)]) %*%((u[i]-u[(i-1):(i-2)]%*%ml)/(df+(u[i]-u[(i-1):(i-2)]%*%ml)^2))

################ LS and MLE truncated by LS

ls=ls+im%*% as.matrix(u[(i-1):(i-2)]) %*%(u[i]-u[(i-1):(i-2)]%*%ls)

tml2=tml2+(df+3)*im1 %*% as.matrix(u[(i-1):(i-2)]) %*%((u[i]-u[(i-1):(i-2)]%*%tml2)/(df+(u[i]-u[(i-1):(i-2)]%*%tml2)^2))
```

```
if (norm(tml2-ls,"f")>C*i^(e)){tml2=ls+C*i^(e)*(tml2-ls)/norm(tml2-ls,"f");t=i} ## Truncated by
Eular distance



################## Store the current estimators
LSE=cbind(LSE,ls)
MLT=cbind(MLT,tml)
MLE=cbind(MLE,ml)
MLTLS=cbind(MLTLS,tml2)

}############################################## End inner loop

T=T+t                                   ## Add up index No. of the latest step where truncation works

######################### Update Mean Square Errors
MSEmlt= MSEmlt*(j-1)/j +(((MLT-theta)[1,])^2+((MLT-theta)[2,])^2)/j
MSEmle= MSEmle*(j-1)/j +(((MLE-theta)[1,])^2+((MLE-theta)[2,])^2)/j
MSElse= MSElse*(j-1)/j +(((LSE-theta)[1,])^2+((LSE-theta)[2,])^2)/j
MSEmltls= MSEmltls*(j-1)/j +(((MLTLS-theta)[1,])^2+((MLTLS-theta)[2,])^2)/j

######################### Update Means
MEANmlt= MEANmlt*(j-1)/j +MLT/j
MEANmle= MEANmle*(j-1)/j +MLE/j
MEANlse= MEANlse*(j-1)/j +LSE/j
MEANmltls= MEANmltls*(j-1)/j +MLTLS/j

}################## end outer loop
par(mfrow=c(1,1))
if(Plots==3){par(mfrow=c(1,2))}

########## Plotting the Mean
if(Plots==1|Plots==3){
plot(MEANlse[1,],,'l' ,
xlab="Observations",ylab="")                    ######### LS Lines
lines(MEANlse[2,])
lines(rep(t1,(n-1)),lty=2)
lines(rep(t2,(n-1)),lty=2)
lines(MEANmle[1,],lty=1,col="blue")             ####### MLE Dotted Lines
lines(MEANmle[2,],lty=1,col="blue")
lines(MEANmltls[1,],lty=1,col="red")            ####### MLT Dotted Lines
lines(MEANmltls[2,],lty=1,col="red")
lines(MEANmlt[1,],lty=1,col="yellow")           ####### MLTLS Dotted Lines
lines(MEANmlt[2,],lty=1,col="yellow")
legend("topright", col = c("black", "blue", "yellow", "red"),
legend = c("RLS", "RML", "RMLtri", "RMLls"),
lty = 1, merge = TRUE)
}

########## Plotting MSE
if(Plots==2|Plots==3){
plot(MSElse[1,],,'l',xlab="Observations",ylab="")      ######### RLS black lines
lines(MSEmle[1,],lty=1,col="blue")                     ####### RML blue Lines
lines(MSEmlt[1,],lty=1,col="yellow")                    ####### RMLtri yellow Lines
lines(MSEmltls[1,],lty=1,col="red")                     ####### RMLls red Lines
legend("topright", col = c("black", "blue", "yellow", "red"),
legend = c("RLS", "RML", "RMLtri", "RMLls"),
lty = 1, merge = TRUE)
}
```

```
########## Plotting MSE in later stage
if(Plots==4){
plot(c((n-105):(n-5)),MSElse[1,(n-105):(n-5)],'l',xlab="Observations",ylab="",ylim=c(0.002,0.006))
lines(MSEmle[1,1:(n-5)],lty=1,col="blue")            ####### MLE blue Lines
lines(MSEmlt[1,1:(n-5)],lty=1,col="yellow")          ####### MLT yellow Lines
lines(MSEmltls[1,1:(n-5)],lty=1,col="red")           ####### MLTLS red Lines
legend("topright", col = c("black", "blue", "yellow", "red"),
legend = c("RLS", "RML", "RMLtri", "RMLls"),
lty = 1, merge = TRUE)
}
```

T/R

Bibliography

[1] Anderson, B., and Moore, J. B. A matrix kronecker lemma. *Linear Algebra and its Applications 15*, 3 (1976), 227–234.

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

[3] Andrieu, C., Moulines, É., and Priouret, P. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on control and optimization 44*, 1 (2005), 283–312.

[4] Benaim, M. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization 34*, 2 (1996), 437–472.

[5] Benaïm, M., and Hirsch, M. W. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations 8*, 1 (1996), 141–176.

[6] Benaïm, M., and Hirsch, M. W. Stochastic approximation algorithms with constant step size whose average is cooperative. *Annals of Applied Probability* (1999), 216–241.

[7] Benaïm, M., Hofbauer, J., and Sorin, S. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization 44*, 1 (2005), 328–348.

[8] Benveniste, A., Métivier, M., and Priouret, P. *Adaptive algorithms and stochastic approximations*. Springer Publishing Company, Incorporated, 2012.

[9] Blum, J. R. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics* (1954), 737–744.

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

[11] Bray, M. M., and Savin, N. E. Rational expectations equilibria, learning, and model specification. *Econometrica: Journal of the Econometric Society* (1986), 1129–1160.

[12] Burkholder, D. L. On a class of stochastic approximation processes. *The Annals of Mathematical Statistics* (1956), 1044–1059.

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

[14] Chen, H. *Stochastic approximation and its applications*, vol. 64. Springer Science & Business Media, 2002.

[15] 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.

[16] 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.

[17] Chung, K. L. On a stochastic approximation method. *The Annals of Mathematical Statistics* (1954), 463–483.

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

[19] Delyon, B., Lavielle, M., and Moulines, E. Convergence of a stochastic approximation version of the em algorithm. *Annals of Statistics* (1999), 94–128.

[20] Dvoretzky, A., et al. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (1956), The Regents of the University of California.

[21] 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.

[22] Evans, G. W., and Honkapohja, S. Local convergence of recursive learning to steady states and cycles in stochastic nonlinear models. *Econometrica: Journal of the Econometric Society* (1995), 195–206.

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

[24] Feigin, P. D. Conditional exponential families and a representation theorem for asympotic inference. *The Annals of Statistics* (1981), 597–603.

[25] Gladyshev, E. G. On stochastic approximation. *Theory of Probability & Its Applications 10*, 2 (1965), 275–278.

[26] Gu, M. G., and Li, S. A stochastic approximation algorithm for maximum-likelihood estimation with incomplete data. *Canadian Journal of Statistics 26*, 4 (1998), 567–582.

[27] Hall, P., and Heyde, C. Martingale limit theory and applications. *Academic, New York* (1980).

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

[29] Haykin, S. Neural networks: a comprehensive foundation. *Mc Millan, New York* (1994).

[30] Heyde, C. C. *Quasi-likelihood and its application: a general approach to optimal parameter estimation*. Springer Science & Business Media, 2008.

[31] Hodges, J. L., Lehmann, E. L., et al. Two approximations to the robbins-monro process. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (1956), The Regents of the University of California.

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

[33] Huber, P. J. Robust statistics, 1981.

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

[35] Kiefer, J., Wolfowitz, J., et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics 23*, 3 (1952), 462–466.

[36] Kushner, H., and Clark, D. Stochastic approximation for constrained and unconstrained systems, volume 26 of applied mathematical sciences, 1978.

[37] Kushner, H. J. General convergence results for stochastic approximations via weak convergence theory. *Journal of mathematical analysis and applications 61*, 2 (1977), 490–503.

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

[39] Kushner, H. J., and Shwartz, A. An invariant measure approach to the convergence of stochastic approximations with state dependent noise. *SIAM Journal on Control and Optimization 22*, 1 (1984), 13–27.

[40] Kushner, H. J., and Yin, G. *Stochastic approximation algorithms and applications*. Springer, 1997.

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

[42] Lai, T. L., and Robbins, H. Adaptive design and stochastic approximation. *The annals of Statistics* (1979), 1196–1221.

[43] Lai, T. L., and Robbins, H. Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. *Probability Theory and Related Fields 56*, 3 (1981), 329–360.

[44] Lai, T. L., and Wei, C. Z. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics* (1982), 154–166.

[45] Lai, T. L., and Ying, Z. Efficient recursive estimation and adaptive control in stochastic regression and armax models. *Statistica Sinica 16*, 3 (2006), 741.

[46] 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.

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

[48] Lehmann, E. L., and Casella, G. *Theory of point estimation*, vol. 31. Springer, 1998.

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

[50] Ljang, L., and Soderstrom, T. Theory and practice of recursive identification, 1987.

[51] Ljung, L. Analysis of recursive stochastic algorithms. *Automatic Control, IEEE Transactions on 22*, 4 (1977), 551–575.

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

[53] Masreliez, C., and Martin, R. Robust bayesian estimation for the linear model and robustifying the kalman filter. *Automatic Control, IEEE Transactions on 22*, 3 (1977), 361–371.

[54] Nevelson, and Khasminskiĭ. *Stochastic approximation and recursive estimation.*

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

[56] Polyak, B. T., and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization 30*, 4 (1992), 838–855.

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

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

[59] Robbins, H., and Siegmund, D. A convergence theorem for nonnegative almost supermartingales and some applications. *in JS Rustagi, ed., (Optimizing Methods in Statistics) Academic Press, New York* (1971), 233–257.

[60] 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.

[61] Sacks, J. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics* (1958), 373–405.

[62] 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.

[63] Serfling, R. Approximation theorems of mathematical statistics, 1980.

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

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

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

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

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

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

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

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

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

[73] 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.

[74] Watkins, C., and Dayan, P. Q-learning. *Machine learning 8*, 3-4 (1992), 279–292.

[75] Wei, C. Z. Multivariate adaptive stochastic approximation. *The Annals of Statistics* (1987), 1115–1130.

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