Outer Products and Stochastic Approximation Algorithms in a Heavy-tailed and Long-range Dependent Setting

by

Samira Sadeghi

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

As has been established in the literature, the rate of convergence in the strong law of large numbers for a centered stationary and asymptotically independent time series $(X_k)_{k\in\mathbb{Z}}$ with finite moment of order $p \in [1, 2)$, is given by any integer $q < p$ such that $n^{-\frac{1}{q}} \sum_{i=1}^{n} X_i$ converges almost surely to 0. This type of result is called a Marcinkiewicz strong law of large number. If the tails of probability distribution X_i is heavier than the exponential distribution then X_i is heavy-tailed.

Based on the classical definition of long-range dependence, a time series is declared long-range dependent if the sum of the autocovariances diverges. Hall [51] suggested that the long-range dependence should be considered in view of a specific convergence problem, and a time series should be considered long-range dependent if the convergence rate in the problem of interest is strictly slower than in the case of independent data.

Classical time-series theories are mainly concerned with the statistical analysis of light-tailed and short-range dependent stationary linear processes. Applications in network theory and financial mathematics lead us to consider time series models with heavy tails and long memory. Heavy-tailed data exhibits frequent extremes and infinite variance, while positively-correlated long memory data displays great serial momentum or inertia. Heavy-tailed data with long-range dependence has been observed in a plethora of empirical data set over the last fifty years and so. Methodological and theoretical results as well as a considerable portion of applied work in this thesis address long-range dependence and heavy-tailed types of the data.

The first contribution of this thesis is the development of Marcinkiewicz strong law of large numbers for outer products of multivariate linear processes while handling long-range dependent and heavy-tailed data structure. This result is used to obtain Marcinkiewicz strong law of large numbers for non-linear function of partial sums, sample auto-covariances and linear processes in a stochastic approximation setting.

The next part of the result is on developing almost sure convergence rates for linear stochastic approximation algorithms under some assumptions that are implied by Marcinkiewicz strong law of large numbers. Finally, we verify our results experimentally in the stochastic approximation setting while handling all gains, long-range dependence and heavy tails and addressing the optimal polynomial rate of convergence by establishing results akin to the Marcinkiewicz strong law of large numbers.

Preface

Chapter 2 and 3 of this thesis are based on an accepted paper and a published paper as:

. Kouritzin, M.A. and Sadeghi, S. (2015). *Marcinkiewicz Law of Large Numbers for Outer-products of Heavy-tailed, Long-range-Dependence Data.* Advances in Applied Probability Journal, in press.

. Kouritzin, M.A. and Sadeghi, S. (2015). *Convergence Rates and Decoupling in Linear Stochastic Approximation Algorithms.* SIAM Journal on Control and Optimization, **vol.** 53-3, pp. 1484-1508.

I was responsible for mathematical analysis, developing the numerical technique and implementing it in a computer code as well as manuscript composition. Kouritzin, M.A. was involved with concept formation, mathematical analysis and manuscript edits. I was the corresponding author in both papers.

To my beloved parents : Hossein and Noorbanoo

and

my twin sister: Samaneh

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Michael A. Kouritzin. Without his guidance, constant encouragement and support, this thesis could never have been possible. It has been a rewarding opportunity for me to work under his supervision. He had a great impact on my thought processes.

I would like to take this opportunity to extend my gratitude to Dr. Lewis and Dr. Frei for being part of my supervisory committee and to Dr. Schmuland and Dr. Horváth for being my examining committee. I thank you all for dedicating your time to reviewing this thesis and providing constructive and insightful comments.

I would like to thank my parents (Hossein and Noorbanoo) and my family for their unending love and unconditional support throughout my life. They have always been there for me when I needed them and have been instrumental in every success of my life.

My sincere thanks goes also to my sister, Samaneh, and all my friends especially Sahar, Mohammad and Meisam.

I am also heartily grateful to Dr. Bry for all his help during the difficult first year of my PhD program. I feel so fortunate to learn from his great personality and invaluable expertise.

Table of Contents

List of Tables

Chapter 1

Introduction

In this chapter, we give background for Marcinkiewicz strong law of large numbers, long-range dependence and heavy tail phenomena as well as stochastic approximation type of algorithms. We finish this chapter by providing the research objectives and notation list. Results given here are widely known; the theorems are stated without proofs.

1.1 Strong Laws under Heavy Tails and Long-Range Dependence

We start with law of large numbers and its evolution to Marcinkiewicz strong law of large numbers. Then we talk about notion of heavy tails and long-range dependence. Finally, we give a literature review on Marcinkiewicz strong law of large numbers for processes with heavy tails and long-range dependence.

1.1.1 From Strong Law of Large Numbers to Marcinkiewicz Strong Law of Large Numbers

The weak law of large number was proved by Swiss mathematician James Bernoulli around 1700 which was published in his treatise *“Ars Conjectandi”* [5] posthumously in 1713. Bernoulli’s theorem was generalized by Poisson [57] around 1800 and Chebychev [11] developed the method under his name in 1866. Later on, Chebychev’s argument was employed to extend Bernoulli’s theorem to dependent random variables by Markov [47]. Further generalization of Bernoulli’s theorem as the strong law of large numbers was proved by French mathematician Emile Borel [7] in 1909. Necessary and sufficient conditions for a set of mutually independent random variables to follow the law of large numbers was derived by Kolmogorov in 1926.

The weak law due to Bernoulli states that

$$P\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right) \leq \frac{pq}{n\epsilon^2}, \tag{1.1}$$

where X_i are independent and identically distributed (i.i.d.) Bernoulli random variables such that $P(X_i = 1) = p, P(X_i = 0) = 1 - p = q$ and $S_n = X_1 + \cdots + X_n$ represent the number of successes in n trials. That is to say, the ratio total number of successes to the total number of trials tends to p in probability as n increases. A stronger version of this result due to Borel and Cantelli is called strong law of large numbers and states that the ratio $\frac{S_n}{n}$ tends to p not only in probability, but also with probability 1. Technically:

If $\{\epsilon_n\}$ is a sequence of positive numbers converging to zero, then the strong law of large numbers states that if

$$\sum_{n=1}^{\infty} P\left(\left|\frac{S_n}{n} - p\right| > \epsilon_n\right) < \infty \tag{1.2}$$

is satisfied then the Borel-Cantelli lemma guarantees that the events of form $\left\{\left|\frac{S_n}{n} - p\right| > \epsilon_n\right\}$ occur only for a finite number of n in an infinite sequence, which means the event $\frac{S_n}{n}$ converges to p almost-surely.

Note that the strong law of large numbers would immediately follow from the weak law of large numbers if it were not for non-summability of $\frac{1}{n}$ in equation (1.1).

Based on the weak law, the ratio $\frac{S_n}{n}$ for all large enough n, is likely to stay close to p with a probability that tends to 1 as n increases. However, if additional trials are conducted beyond specific $n^{'}$ that already satisfied (1.1), the weak law does not guarantee that $\frac{S_n}{n}$ is bounded to stay close p. Indeed, events may occur such that for $n > n^{'}$, $\frac{S_n}{n}$ be greater than $p+\epsilon$. The probability for such an event is the sum of a large number of very small probabilities, and the weak law is unable to say anything specific about the convergence of that sum.

However, the strong law based on (1.2) states that not only all such sums converge, but the total number of all such events where $\frac{S_n}{n}$ is greater than $p + \epsilon$ is indeed finite. This implies that the probability $\left\{\left|\frac{S_n}{n} - p\right| > \epsilon\right\}$ of the events as n increases becomes and remains small, since with probability 1 only finitely many events violate the above inequality as n goes to infinity.

What was defined above is the strong law of large number in the context of binomial distributions. There is a natural analogue to this law regardless of type of probability distributions, which we state as follows.

Theorem 1.1 *Suppose $\{X_i, i \geq 1\}$ be a sequence of i.i.d. random variables with finite expected value $EX_1 = \mu$ and $E(|X_1|) < \infty$. Let $S_n = X_1 + \cdots + X_n$ be the sum of the X_i. Then*

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1 \tag{1.3}$$

The strong law of large number states that the sample average converges almost surely to the expected value.

Taking into consideration that the strong law of large number can be viewed as a result about the magnitude of the fluctuations of $\{S_n, n \geq 1\}$ when $E|X_1| < \infty$, it seems natural to investigate whether there are analogous fluctuation results when $E|X_1|^p < \infty$ for some $1 < p < 2$. This leads us to the generalization of Kolmogorov's strong law of large numbers which is called Marcinkiewicz strong law of large numbers.

Theorem 1.2 *Let $\{X_i, i \geq 1\}$ be a sequence of i.i.d. random variables with finite expected value $EX_1 = \mu$. If $E|X_1|^p < \infty$ for some $1 < p < 2$, then*

$$\frac{(S_n - n\mu)}{n^{\frac{1}{p}}} \to 0 \quad a.s. \tag{1.4}$$

Theorem 1.2 states that the magnitude of the asymptotic fluctuations of S_n about the line nEX_1 are asymptotically of no larger than $n^{\frac{1}{p}}$ when $E|X_1|^p < \infty$ for some $1 < p < 2$.

The classical Kolmogorov's strong law of large numbers has been extended to various weakly dependent random variables which are not necessarily identically distributed. There has been a huge amount of work concerning the rates of convergence in the strong law of large numbers for weakly dependent random variables. Chandra and Ghosal [10] showed that the independence assumptions can be weakened for Marcinkiewicz strong law of large numbers as well.

Bo [6] studied the strong limit behavior of sequences of blockwise m-dependent random variables with only p-moments, where $1 \leq p < 2$, and provide a result analogous to the classical Marcinkiewicz-Zygmund strong law of large numbers which states that if $\{X_n\}_{n\geq 1}$ is a sequence of i.i.d. random variables with $E|X_1|^p < \infty$ and $0 < p < 2$, then

$$\frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (X_k - c) \to 0 \quad a.s. \tag{1.5}$$

where $c = 0$ if $0 < p < 1$ and $EX_1 = c$ if $1 \leq p < 2$.

Definition 1.1 *A sequence of random variables $\{X_1, X_2, \cdots\}$ is m-dependent for some $m \geq 0$, if the vector $(X_1, \cdots, X_i)$ is independent of $(X_{i+j}, X_{i+j+1}, \cdots)$ whenever $j > m$.*

For strongly mixing sequences, the problem of rate of convergence in Marcinkiewicz strong law of large numbers has been fully considered in Rio [63].

Definition 1.2 *The sequence of random variables $\{X_n\}_{n\geq 1}$ is said to be strong mixing if $\alpha(m) \to 0$ as $m \to \infty$ where:*

$$\alpha(m) = \sup_{n\geq 1} \sup_{F\in\mathcal{F}_1^n, G\in\mathcal{G}_{n+m}^{\infty}} |P(F \cap G) - P(F)P(G)|$$

for all $m = 1, 2, \cdots$.

Louhichi [42] proved that the Marcinkiewicz strong law of large numbers holds for associated sequences for which the second moment is not assumed to be finite but rather the moment of order $p(p \in [1, 2[)$. Hence, similar to Dabrowski and Jakubowski [14], she found a suitable weak dependence coefficient defined for associated sequences with infinite variance.

Note that based on one of the given definition of heaviness of tail in the next two sections, Louhichi [42] considered Marcinkiewicz strong law of large numbers for heavy tailed laws.

1.1.2 Limit Theory for Linear Processes

As we mentioned quickly in the previous section the strong laws of large numbers and other variants of limit theorems for independent and identically distributed random variables have been extensively studied and many deep results have been obtained. With various weak dependence conditions, some of the obtained results under the i.i.d. assumption have been generalized to dependent random variables. See e.g., [36], [70], [30], [40], [21], [78] and [23] among others.

The limiting behavior for a sequence of random variables is interesting in its own right. However, it is of interest to see under what conditions the classical limit theories for random variables hold for some linear processes, for example, of form $x_k = \sum_{j=0}^{\infty} c_j \xi_{k-j}$ with coefficients c_j and i.i.d. innovations $\xi_k, k \in \mathbb{Z}$ with zero mean.

Phillips and Solo [59] used an algebraic decomposition method and established (among several other weak and strong results) that under innovations with finite second moment and $\sum_{j=1}^{\infty} j^2 c_j^2 < \infty$, the strong law of large numbers for partial sum of linear processes follows directly from strong law of large numbers of innovations. They also proved that the strong law of large numbers still holds if we relax the second moment boundedness of innovations but instead impose the following condition on the coefficients: $\sum_{j=1}^{\infty} j|c_j| < \infty$.

Therefore, based on adopted conditions on the innovations and coefficients different limit laws can be transferred from the innovations to the corresponding linear processes.

In addition, the limit behavior of dependent linear processes has been investigated in many directions and researchers have established many results concerning strongly mixing processes of various types; see e.g., [4], [8], [24], [69], [63], [40], [18], [60] and [61] among others. Strong mixing is one of the most general and well known type of mixing.

Also, there has been a lot of research concerning the properties of dependent linear processes. For instance, Chanda [9] has shown that members of the important class of linear stochastic processes are strongly strong mixing, provided they are based on innovation random variables which have Lebesgue-integrable characteristic functions. Later, Withers [76] gave an alternative set of conditions for linear processes to be strong mixing. Mokkadem [54] showed that stationary vector ARMA processes are strong mixing, provided the innovations have absolutely continuous distribution with respect to Lebesgue measure. On the other hand, Ibragimov and Linnik [36] and Chernick [12] gave examples of first-order autoregressive (AR(1)) processes based on discrete-valued innovation random variables which are not strong mixing. The latter proof was based on contradiction method and did not give the intuitive reason about what causes the failure of the strong mixing condition. Several years later, Andrews [2] explicitly constructed sequences of sets which violated the strong mixing condition and showed that certain AR(1) processes are not strong mixing. The intuition behind Andrews' construction, as mentioned in

[2], is as follows:

Suppose $\{X_t\}$ is an AR(1) process based on Bernoulli (q) innovation random variables, and $X_{t,s}$ is equal to X_{t+s} minus its component which depends on $X_t, X_{t-1}, \cdots$. If we know X_t is small, then we know that with probability 1 X_{t+s} must fall in a set which is a small neighborhood of the support of $X_{t,s}$. A sequence of such small neighborhoods can be constructed for $s = 1, 2, \cdots$ which have unconditional probability bounded away from 1. Hence, knowledge that X_t is small increases the probability of certain sets which are determined by the "future" of the process, no matter how far in the future, by a non-negligible amount. This implies $\{X_t\}$ is non-strong mixing.

This thesis is not concerned with sum of random variables produced by linear processes, but rather interested in sum of outer products of linear processes which will be discussed more in future sections.

1.1.3 The Notion of Heavy Tails and Long-Range Dependence

Heavy Tails

To detect heavy tails and measuring heaviness of the tails, there are several statistical approaches. Whenever in literature random variables with heavy tails are mentioned, it can be interpreted in different ways [67]. Some common possibilities:

- random variables with subexponential tails
- random variables with regularly varying right tails
- random variables with infinite second moment

It is, therefore, important, to ascertain in what sense the notion of heavy tails is used in any given instance.

Definition 1.3 *The random variable X is called subexponential random variable if*

$$\lim_{x\to\infty} \frac{P(X_1 + X_2 > x)}{P(X > x)} = 2. \tag{1.6}$$

When X_1 and X_2 are two independent copies of X.

A function $f : \mathbb{R} \to \mathbb{R}$ is called regularly varying at infinity with exponent $a \in \mathbb{R}$ if for every $c > 0$

$$\lim_{x\to\infty} \frac{f(cx)}{f(x)} = c^a. \tag{1.7}$$

The function f is called slowly varying (at infinity) if (1.7) holds with $a = 0$. Any function f that is regularly varying with exponent a can be written in the form $f(x) = x^a L(x)$ for $x > 0$, where L is a slowly varying function.

Definition 1.4 *A random variable X is said to have regularly varying right tail with tail exponent $\alpha > 0$, if $P(X > x)$ is regularly varying at infinity with exponent $-\alpha$.*

$$P(X > x) = x^{-\alpha} L(x)$$

If $P(X > x)$ is slowly varying at infinity, then we say that X has a slowly varying right tail.

Generally, heavy tails are related to the tendency of various time series to exhibit sudden and discontinuous changes. This was first observed by Mandelbrot [46] who proposed the stable Paretian distribution for modeling financial time series. The empirical work by Mandelbrot [46] and others lead to the general acceptance that there are heavy tails in financial distributions. The reason for heaviness of tail in financial data is that large observations have non-negligible probability and they, although rare, can dominate a systems performance.

The assumption that the random variables under investigation follow a Gaussian distribution is the basis of many techniques. However, the marginal distributions of observed time series in many areas are heavy-tailed or asymmetric and often deviate from the Gaussian model. In such circumstances, the appropriateness of the commonly adopted normal assumption is highly questionable. Hence, in the presence of heavy tails it is natural to assume that they are approximately controlled by a non-Gaussian stable distribution. Investigations of the appropriateness of the stable Paretian distribution in modeling heavy tail type of data have been started by the work of Fama [25][26] (See, [48], [50], [51], [52], [53], [58] and [62]).

Based on Rachev and Mittnik [52], a random variable X is heavy-tailed distributed with index α if $P(X \geq x) \sim cx^{-\alpha}L(x)$ as $x \to \infty$ for $c > 0$ and $0 < \alpha < 2$ when $L(x)$ is a slowly varying function. Mandelbrot refers to this effect as the infinite variance syndrome which shows observations of a heavy-tailed distribution can fluctuate far from its mean value (defined only when $1 < \alpha < 2$) with non-negligible probability.

Long-Range Dependence

Around the World War II a huge impetus was given to research in time series, as a natural result of developments in such areas as radio signals and some engineering applications. Subsequently, a flexible subset of models, so-called ARMA, was reformulated in the time domain. These are short-range dependent models involving correlation functions that decrease exponentially fast over time. Despite the fact that short-memory models had a wide range of usage, by economists for example, these type of models had certain shortcomings and were not applicable to all fields. For instance, the measurement on the Nile River taken by Hurst [34] [35] in the 1950s appeared to require models, whose correlation functions would decay much more slowly.

In the sixties, models involving the so-called "Joseph effect" or long-range dependence was suggested by Benoit Mandelbrot [44]. Long-range dependence or long memory denotes the property of time series to exhibit persistent behavior such as a significant dependence between very distant observations. generally, Long-range dependence affects phenomena in which correlations decay like a power law, thus much less quickly than in the ARMA models. Mandelbrot generated long-range dependence through the fractional Brownian motion model and its increments. Fractional Brownian motion was discovered by Kolmogorov [38] in 1940 but it was Mandelbrot [45] who distinguished its relevance to applications. Authors like Samorodnitsky, Yaglom, Rosenblatt, Major, Dobrushin, Taqqu, Giraitis, Surgalis, Robinson (see e.g. [66], [79], [65], [20], [72], [27] [64]) and many others continued Kolmogorov's work in the theoretical developments of long-range dependence.

The phenomenon of long-range dependence is widely believed to be both ubiquitous and important in data arising in a variety of different fields, such as econometrics [64], hydrology [56], climate studies [74], DNA sequencing [37] etc. Yet what is long-range dependence? How does one measure it? There is no consensus on the notion of heavy tails. There is even less consensus on the notion of long-range dependence. Historically, long-range dependence was viewed as a property of certain stochastic models with a finite variance, and then it was associated either with a particularly slow decay of correlation. Long-range dependence is a phenomenon characterized by sample paths displaying apparent trends and cycles. Processes whose autocorrelation function, decaying as a power law in the lag variable for large lag values, sums to infinity. The decay is slower than exponential, and the area under the curve is infinite. One of the definition of long-range dependence is as follows:

Definition 1.5 *A stationary process $\{X_k\}$ (with finite variance) is said to have long-range dependence if its autocorrelation function $\rho(h) = corr(X_k X_{k+h})$ decays as a power of the lag h:*

$$\rho(h) = corr(X_k X_{k+h}) \sim \frac{L(h)}{h^{1-2\sigma}} \qquad as\, h \to \infty,\ 0 < \sigma < \frac{1}{2} \tag{1.8}$$

where L is slowly varying at infinity.

In other word, if the sum of the autocovariances diverges then the series is said long-range dependent. Concentrating too much on the correlations, however, has a number of drawbacks. For example, correlations provide only very limited information about the process if the process is not very close to being Gaussian. In addition to obvious drawbacks of correlation carrying limited information in non-Gaussian case, this leaves one unable to define long memory for stochastic processes with infinite variance.

Models in which variances may be infinite are needed for the analysis of a variety of phenomena (e.g. in finance [49], geology [55]). The above definition of long-range dependence is inflexible in providing extensions that allow for departure from stationarity and, indeed, infinite variances.

The problem of defining long-range dependence for infinite variance time series is made even more ambiguous because of the fact that there is not a unique structure that can describe such time series. Heyde and Yang [31] provided definitions, almost equivalent to the original ones in that domain of applicability, which were useful for processes which may not be second-order stationary, or indeed have infinite variances.

Samorodnitsky [67] proposed a new way of thinking about long-range dependence in terms of the way rare events happen. This is particularly appropriate in the heavy-tailed situations because most practitioners using heavy-tailed models are interested precisely in certain rare events related to the tails. However, since this thesis is not concerned about all different definitions of long-range dependence, we end the discussion here.

In the next section we define the notion of long-range dependence and heavy tailness that is used through the rest of this thesis.

1.1.4 Marcinkiewicz Strong Law of Large Numbers with Heavy-Tails and/or Long-Range Dependence

Let $\{X_k\}$ be $\mathbb{R}^d$-valued (possibly two-sided, multivariate) linear processes

$$X_k = \sum_{l=-\infty}^{\infty} C_{k-l}\Xi_l, \tag{1.9}$$

defined on some probability space (Ω, F, P) with coefficient matrix (C_l) and i.i.d. zero-mean innovations $\{\Xi_l\}$. If the coefficients (C_l) are absolutely summable and innovations have second moments, then the covariances of X_k are summable and we say that $\{X_k\}$ is short-range dependence. On the contrary, we generically say that $\{X_k\}$ is long-range dependence if its covariances are not absolutely summable. Practically, as $|l| \to \infty$ by choosing appropriate coefficients, matrix sequence (C_l) can decay slowly enough such that $\{X_k\}$ shows long-range dependence. We consider $\{X_k\}$ to have long-range dependence too in this $\{C_l\}$ non-summable case even though the second moments for X_k may not exist. If each X_k fails to have a second moment, then we say it is heavy-tailed (HT) and is otherwise light-tailed (LT).

The limit behavior of linear processes with heavy-tailed and/or long-range dependence has been investigated for a long time in different contexts such as partial sums, sample covariance and non-linear function of partial sums. Though, there are limited number of works on Marcinkiewicz strong law of large numbers for partial sums of X_k under both heavy-tailed and the long-range dependence and there is hardly any result handling the Marcinkiewicz strong law of large numbers for partial sums of nonlinear functions of X_k.

It is widely known that if $(X_n)_{n\in\mathbb{Z}}$ is an i.i.d. sequence of random variables in the domain of attraction of stable law with index α, $1 < \alpha \leq 2$, then $n^{-\frac{1}{\alpha}} \sum_{i=1}^{n} X_i$ converges in distribution to a stable law, however, for all $p < \alpha$ the statistic $n^{-\frac{1}{p}} \sum_{i=1}^{n} X_i$ converges almost surely to zero. This problem has been investigated for weakly dependent variables as well, see for instance Rio [63] and the references therein, but has not been fully studied in the context of long-range dependence. One result in this area is by Louhchi and Soulier [43] in which they considered the Marcinkiewicz strong law of large numbers for associated random variables with not necessarily finite second moment.

Nonetheless, there have been large number of studies concerning the weak convergence of linear processes. For instance, many results handle the existence and description of limit distributions of sums

$$S_{n,h}(t) = \sum_{k=1}^{[nt]} (h(x_k) - E(h(x_k))), \quad t \geq 0, \tag{1.10}$$

where h is a (nonlinear) function. One-sided linear (moving average) process is one of the well studied non-Gaussian long-range dependent processes,

$$x_k = \sum_{j=0}^{\infty} c_j \xi_{k-j}, \tag{1.11}$$

in which, innovations $\xi_k, k \in \mathbb{Z}$, are i.i.d., have zero mean with finite variance, and coefficients c_j satisfy:

$$c_j \sim c_\sigma j^{-\sigma}, \quad j \geq 1 \tag{1.12}$$

for some constant $c_\sigma \neq 0$, $c_0 = 1$ and $\sigma \in (\frac{1}{2}, 1)$.

The distributional convergence for normalized partial sums of Appell polynomials $A_m(x_k)$ of linear processes x_k having both long-memory and heavy tails in the sense $EA_m^2(x_k) = \infty$ has been studied in Vaiciulis [73]. In particular, he assumed x_k had the form (1.11) with innovations belonging to the domain of attraction of an α-stable law with $1 < \alpha < 2$ and c_j following (1.12).

Another example which, in the case of long-range dependent and heavy tails, is an area of active research, involves auto-covariance functions. The research in this area started with work of Anderson and Walker [1], which resulted in a central limit theorem for the so-called autocorrelation process under strict stationarity and summability condition of coefficients. In fact, Anderson and Walker showed that the normal approximation can continue to hold even when ξ_k does not have finite fourth-order moments but, as illustrated in [15], this is limited to one-dimensional autocorrelation processes. Later, Hannan [32] added the autocovariance process to the class of strictly stationary, linear-model-based processes satisfying the central limit theorem under general conditions including finite fourth-order moments for ξ_k. Next, in studying the sample covariance process, Giraitis and Surgailis [28] proved a central limit theorem for the related process and allowed both processes to be two-sided linear processes. In a complimentary and very interesting set of results, Davis and Resnick [16][17] have established weak convergence results for sample covariance processes of two-sided linear models to non-normal stable distributions when ξ_k does not have fourth order moments with absolutely summable coefficients c_j with form of (1.12). Later, Horváth and Kokoszka [33] considered the asymptotic distribution of normalized sample autocovariances of long-memory processes with innovations of infinite fourth moment.

1.2 Stochastic Approximation Algorithms

Stochastic approximation methods are a family of iterative stochastic optimization algorithms that attempt to find zeroes or extrema of functions which cannot be computed directly, but only estimated via noisy observations (Y). The original work in recursive stochastic algorithms was by the Robbins and Monro, who developed and analyzed recursive procedure for finding the root of real-valued function g and real variable h. Suppose that we have a function $g(h)$ and a constant r, such that the equation $g(h) = r$ has a unique root at $h = c$. While the function $g(h)$ cannot be observed directly but measurements can instead be obtained of the random variable Y where $E[Y(h)] = g(h)$.

Robbins and Monro proved that the following iterative algorithm of the form:

$$h_n = h_{n-1} + \mu_n(r - Y(h_n))$$

generates iterates which is convergent in L^2 to c under some proper conditions.

Nevertheless, as explained in [19], the first known stochastic algorithm has been traced back to 1890 by B. Bru in the European artillery regulations. The problem was to adjust the slant h (an angle) of a cannon in order to obtain a specified range r. This was done by trial and error, firing one shell after another. The nth experiment was done with the adjustment corresponding to the $(n-1)$th estimated value of h, and led to an improved value h_n; the result of this experiment was an observed range r_n, with a misadjustment $r - r_n$. It had been observed that r_n was actually a random variable, what made the problem much more difficult. After some heuristics, the specialists of the army converged on the following algorithm

$$h_n = h_{n-1} - \frac{\lambda}{n}(r - r_n)$$

where λ is some fixed normalization constant. In the case where one observes only if the shot were too short or too long, the regulations were to use

$$h_n = h_{n-1} - \frac{\lambda}{n} sign(r - r_n)$$

The important discovery here is that the right choice of gain is of order $\frac{1}{n}$. This is a common feature of stochastic algorithms to have a rather large gain due to the necessity to average out the noise coming from the randomness of the observations.

Stochastic approximation creates an adaptive filtering algorithm which can be used to produce sequential estimates of parameters and uses feedback in the form of an error signal to refine its transfer function to match the changing parameters. One of the simplest subclasses of general stochastic approximation algorithms could be considered as linear stochastic approximation algorithms, they have a wide range of applications in system identification, adaptive control, transmission systems, adaptive filtering for signal processing, and several aspects of pattern recognition and learning (see e.g., [3], [22], [41], [68], [77]). The aim of these algorithms is to estimate recursively an unknown time invariant (or slowly varying) parameter vector in a target model. Consequently, their asymptotic rates of almost sure and r^{th}-mean convergence as well as invariance and large deviation principles are of utmost importance.

In the class of linear stochastic approximation, the least mean square algorithm has become one of the most popular adaptive filtering algorithm due to its simplicity and robustness. Suppose

$$y_{k+1} = x_k^T h + \epsilon_k \qquad \forall k = 1, 2, \ldots, \tag{1.13}$$

where $\{x_k, k = 1, 2, \cdots\}$ and $\{y_k, k = 2, 3, \cdots\}$ are second order $\mathbb{R}^d-$ and $\mathbb{R}-$valued stochastic processes, defined on some probability space $(\Omega, \mathcal{F}, P)$, h is an unknown d-dimensional parameter or weight vector of interest and ϵ_k is a noise sequence. One of the questions of interest is finding the value of h that minimizes the mean-square error $h \to E|y_{k+1} - x_k^T h|^2$. Assuming the expectations exist, wide-sense stationarity conditions, the value that minimizes the mean-square error is given by $h = E(x_k x_k^T)^{-1} E(y_{k+1} x_k)$. However, it is the case that neither the joint distribution of (x_k, y_{k+1}) nor the necessary stationarity of the processes are known. An alternative method estimates h through a linear algorithm of the form

$$h_{k+1} = h_k + \mu_k(y_{k+1}x_k - x_k x_k^T h_k), \tag{1.14}$$

where μ_k is the k^{th} step size. A general model that includes the least mean square algorithm, as well as other adaptive-filtering algorithms is

$$h_{k+1} = h_k + \mu_k(b_k - A_k h_k), \tag{1.15}$$

where

$$A_k = \frac{1}{N} \sum_{l=\max\{k-N+1,1\}}^{k} x_l x_l^T \text{ and } b_k = \frac{1}{N} \sum_{l=\max\{k-N+1,1\}}^{k} y_{l+1} x_l \tag{1.16}$$

for some $N \in \mathbb{N}$, are random sequences of symmetric, positive-semi-definite matrices and vectors respectively.

The convergence rate and algorithm effectiveness is influenced by step size μ_k (see, e.g., [12], [15] and references cited therein). In an extreme case of homogeneous, deterministic setting, i.e. $A_k = A$ and $b_k = b$, (1.15) can solve the linear equation $Ah = b$. In this situation the best choice is a constant gain, $\mu_k = \epsilon$, and we get $h_n \to h$ geometrically, provided ϵ is small enough that the eigenvalues of $I - \epsilon A$ are within the unit disc. By way of contrast, in the presence of persistent noise, the decreasing step sizes are required for the convergence $h_n \to h$ to take place. Existing results show that the best possible almost-sure rate of convergence is $|h_n - h| = O\left(\sqrt{n^{-1}\log\log(n)}\right)$, implied by the law of the iterated logarithm, and that this rate is only attainable when $\mu_k = \frac{1}{k}$, second moments of A_k, b_k exist and there is no long-range dependence.

(These claims follow from the almost-sure invariance principle in Kouritzin [39].)

There were many results that gave convergence or rates of convergence for linear algorithms. However, these results assumed a specific dependency structure and, thereby, were not generally applicable. More recently, some authors, e.g. [13], [19] and [71], have followed the similar path of transferring convergence and rates of convergence from partial sums of (the coefficient) random variables to the solutions of linear equations.

1.3 Research Objective

As we mentioned in previous sections, there is almost no general Marcinkiewicz strong law of large numbers results for partial sums of X_k neither for partial sums of nonlinear functions of X_k under both heavy-tailed and the long-range dependence. Our objective in part of this thesis is to establish a method and a structure under which certain Marcinkiewicz strong law of large numbers for heavy-tailed and the long-range dependent phenomena can be handled properly. Precisely, for some $(\sigma, \overline{\sigma}) \in \left(\frac{1}{2}, 1\right]$ and $\alpha > 1$ our goal is to prove the Marcinkiewicz Strong Law,

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (D_k - D) = 0 \quad \text{a.s.} \quad \text{for} \quad p < \frac{1}{2-\sigma-\overline{\sigma}} \wedge \alpha \wedge 2,$$

for outer products $D_k = X_k \overline{X}_k^T$, where $\{X_k\}, \{\overline{X}_k\}$ are both two-sided multivariate linear processes of form

$$X_k = \sum_{l=-\infty}^{\infty} C_{k-l}\Xi_l, \quad \overline{X}_k = \sum_{l=-\infty}^{\infty} \overline{C}_{k-l}\overline{\Xi}_l. \tag{1.17}$$

with i.i.d. zero-mean random $\mathbb{R}^{m+m}$-vectors (innovations) such that $E[|\Xi_1|^2] < \infty$, $E[|\overline{\Xi}_1|^2] < \infty$ and $(C_l)_{l\in\mathbb{Z}}$, $(\overline{C}_l)_{l\in\mathbb{Z}}$ are $\mathbb{R}^{d\times m}$-matrix sequences satisfying $\sup\limits_{l\in\mathbb{Z}} |l|^\sigma \|C_l\| < \infty$, $\sup\limits_{l\in\mathbb{Z}} |l|^{\overline{\sigma}} \|\overline{C}_l\| < \infty$. As $|l| \to \infty$, matrix sequences C_l and $\overline{C}_l$ can decay slowly enough that $\{X_k, \overline{X}_k\}$ have long-range dependence while $\{D_k\}$ can have heavy tails. In particular, the heavy tail and long-range-dependence phenomena for $\{D_k\}$ are handled simultaneously and a new decoupling property is proved that shows the convergence rate is determined by the worst of the heavy tails requirement $p < (\alpha \wedge 2)$ or the long-range dependent condition $p < \frac{1}{2-\sigma-\overline{\sigma}}$, but not the combination.

The main result is applied to obtain Marcinkiewicz strong law of large numbers for non-linear functions forms, autocovariances and stochastic approximation.

Next part of research objective is studying almost sure convergence rates for linear algorithms

$$h_{k+1} = h_k + \frac{1}{k^\chi}(b_k - A_k h_k) \tag{1.18}$$

where $\chi \in (0, 1)$, $\{A_k\}_{k=1}^{\infty}$ are symmetric, positive semidefinite random matrices and $\{b_k\}_{k=1}^{\infty}$ are random vectors. It is shown that $n^\gamma |h_n - A^{-1}b| \to 0$ a.s. for the $\gamma \in [0, \chi)$, positive definite A and vector b such that $\frac{1}{n^{\chi-\gamma}} \sum\limits_{k=1}^{n} (A_k - A) \to 0$ and $\frac{1}{n^{\chi-\gamma}} \sum\limits_{k=1}^{n} (b_k - b) \to 0$ a.s. When $\chi - \gamma \in \left(\frac{1}{2}, 1\right)$, these assumptions are implied by the Marcinkiewicz strong law of large numbers, which allows the

$\{A_k\}$ and $\{b_k\}$ to have heavy tails, long-range dependence or both. The idea is inferring convergence and rates of convergence results for linear algorithms (1.18) from like convergence and rates of convergence of its coefficients.

Finally, we verify our results experimentally in the stochastic approximation setting. In fact, we deal with all gains, long-range dependence, and heavy tails, addressing the optimal polynomial rate of convergence by establishing results akin to the Marcinkiewicz strong law of large numbers, namely $n^{\gamma}|h_n - h| \to 0$ a.s. (i.e. $|h_n - h| = o(n^{-\gamma})$), for all $\gamma < \gamma_0(\chi) \doteq \chi - M$. M is called the *Marcinkiewicz threshold* in what follows and is defined by

$$M \doteq \inf\left\{\frac{1}{m} : \lim_{n\to\infty} \frac{1}{n^{\frac{1}{m}}} \sum_{k=1}^{n} (A_k - A) = 0, \lim_{n\to\infty} \frac{1}{n^{\frac{1}{m}}} \sum_{k=1}^{n} (b_k - b) = 0 \ \text{ a.s.}\right\}. \tag{1.19}$$

Generally, due to strong law of large numbers and central limit theorem in the light-tailed, short-range dependence case, M is expected to vary in the range $M \in (\frac{1}{2}, 1]$, but when there is long-range dependence and/or heavy tails M usually cannot approach $\frac{1}{2}$.

The simulation results show that convergence ($h_k \to h$) in (1.18) takes place provided that $\chi \in (M, 1)$. All this suggests that more quickly decreasing gains like $\mu_k = \frac{1}{k^{\chi}}$ with χ near 1 should be used in very heavy-tailed or long-range dependent settings. Conversely, slowly deceasing gains like $\mu = \frac{1}{k^{\chi}}$ with smaller χ might work well in lighter-tailed, short-range dependent situations. In addition, based on simulations it is clear that the smallest normalized error, $\frac{|h_n - h|}{|h_1 - h|}$, usually occurs for $\chi \in (M, 1]$ and the most commonly used choice $\chi = 1$ is most appropriate in very heavy-tailed or long-range dependent settings (where M is close to 1) or very long runs. In other words, a slower decreasing gain usually gets you close to the true parameters h more quickly unless the coefficients have a high probability of differing significantly from their means.

1.4 Notation List

Before moving to the result chapters, we define our notations that will be used through this thesis.

$\lvert x \rvert$	is Euclidean distance of some $\mathbb{R}^d$-vector x.
$\lVert C \rVert$	is $\sup_{\lvert x \rvert = 1} \lvert Cx \rvert$ for any $\mathbb{R}^{n \times m}$-matrix C.
$\vert\vert\vert \, A \, \vert\vert\vert^2$	$\sum_{n=1}^{d} \sum_{o=1}^{d} (A^{(n,o)})^2$.
$A^{(n,o)}$	is the $(n,o)^{\mathrm{th}}$ components of $A \in \mathbb{R}^{d \times d}$.
$\lfloor t \rfloor$	is $\max\{i \in \mathbb{N}_0 : i \leq t\}$for any $t \geq 0$.
$\lceil t \rceil$	is $\min\{i \in \mathbb{N}_0 : i \geq t\}$ for any $t \geq 0$.
$a_{i,k} \overset{i}{\ll} b_{i,k}$	means $\forall k$, $\exists c_k > 0$ not depending on i s.t. $\lvert a_{i,k} \rvert \leq c_k \lvert b_{i,k} \rvert \; \forall i, k$.
$\prod\limits_{l=p}^{q} B_l$	is $B_q B_{q-1} \cdots B_p$ if $q \geq p$ or I if $p > q, \forall \, B_l \in \mathbb{R}^{d \times d}$.
$a \vee b$	is $\max\{a, b\}$.
$a \wedge b$	is $\min\{a, b\}$.

Bibliography

[1] Anderson, T.W. and Walker, A.M. (1964). *.On the asymptotic distribution of the autocorrelations of a sample from a linear process.* Ann. Math. Statist., **vol.** 35, pp. 1296-1303.

[2] Andrews, D.W.K. (1984). *Non-strong mixing autoregressive processes.* Journal of Applied Probability, **vol.** 21, pp. 930-934.

[3] Benveniste, A., Métivier, M., Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag.

[4] Berkes, I. and Philipp, W. (1979). *Approximation theorems for independent and weakly dependent random vectors.* Ann. Probab., **vol.** 7, pp. 29-54.

[5] Bernoulli, J. (1713). *Ars Conjectandi.* Opus Posthumum, Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis, Basileae, (Ch. 1-4 translated into English by Sung B. (1966). *Ars Conjectandi*, Technical Report No. 2, Dept. of Statistics, Harvard University.

[6] Bo, Z. (1998). *Marcinkiewicz-Zygmund law for sequences of blockwise m-dependent random variables.* Statist.Probab.Letters, **vol.** 38, pp. 83-88.

[7] Borel, E. (1909). *Les probabilites denombrables et leurs applications aritbmttiques.* Rendiconti del Circolo Matematico di Palermo, **vol.** 27, pp. 247-271.

[8] Bradley, R.C. (1983). *Approximation theorems for strongly mixing random variables.* Michigan Math. J., **vol.** 30, pp. 69-81.

[9] Chanda, K.C. (1974). *Strong mixing properties of linear stochastic processes.* Journal of Applied Probability, **vol.** 11, pp. 401-408.

[10] Chandra, T.K. and Ghosal, S. (1996). *Extensions of the Strong Law of Large Numbers of Marcinkiewicz and Zygmund for Dependent Variables.* Acta Math. Hungar., **vol.** 71(4), pp. 327-336.

[11] Chebyshev, P.L. (1867). *Des valuers moyennes.* Journal de Mathématiques Pures et Appliquées, **vol.** 12, pp. 177-184.

[12] Chernick, M.R. (1981). *A limit theorem for the maximum of autoregressive processes with uniform marginal distributions.* Ann. Prob., **vol.** 9, pp. 145-149.

[13] Chong, E.K.P., Wang, I.J. and Kulkarni, S.R. (1999). *Noise conditions for prespecified convergence rates of stochastic approximation algorithms.* IEEE Trans. Inform. Theory, **vol.** 45, pp. 810-814.

[14] Dabrowski, A.R., Jakubowski, A. (1993). *Stable limits for associated random variables.* Ann. Probab., **vol.** 1(22), pp. 1-16.

[15] Davis, R.A. and Marengo, J.E. (1990). *Limit theory for sample covariance and correlation matrix functions of a class of multivariate linear processes.* , **vol.** 6, pp. 483-497.

[16] Davis, R.A. and Resnick, S.I. (1985). *Limit theory for moving averages of random variables with regularly varying tail probabilities.* The Annals of Probability, **vol.** 13, pp. 179-195.

[17] Davis, R.A. and Resnick, S.I. (1986). *Limit theory for the sample covariance and correlation functions of moving averages.* Ann. Statist., **14,** pp. 533-558.

[18] Dedecker, J. and Prieur, C. (2004). *Coupling for τ-dependent sequences and applications.* J. Theoret. Probab., **vol.** 17, pp. 861-885.

[19] Delyon, B. (2000). *Stochastic approximation with decreasing gain: Convergence and asymptotic theory.* Tech. report, Universit de Rennes, Rennes, France.

[20] Dobrushin, R.L. and Major, P. (1979). *Non-central limit theorems for non-linear functions of Gaussian fields.* Z. Wahrscheinlichkeitstheorie Verw. Geb., **vol.** 50, pp. 27-52.

[21] Doukhan, P. (2003). *Models, inequalities, and limit theorems for stationary sequences.* In Theory and Applications of Long-Range Dependence (P. Doukhan, G. Oppenheim and M. S. Taqqu, eds.), Birkhäuser, Boston, pp. 43-100.

[22] Duflo, M. (1996). *Algorithmes Stochastiques.* Springer.

[23] Eberlein, E. and Taqqu, M.S., eds. (1986). *Dependence in Probability and Statistics: A Survey of Recent Results.* Birkhäuser, Boston.

[24] Eberlein, E. (1986). *On strong invariance principles under dependence assumptions.* Ann. Probab., **vol.** 14, pp. 260-270.

[25] Fama, E. (1963). *Mandelbrot and the stable Paretian hypothesis.* Journal of Business, **vol.** 36, pp. 420-429.

[26] Fama, E. (1965). *The behavior of stock market prices.* Journal of Business, **vol.** 38, pp. 34-105.

[27] Giraitis L. and Surgailis, D. (1989). *Limit theorem for polynomials of linear process with long-range dependence.* Lith. Math. J., **vol.** 29, pp. 128-145.

[28] Giraitis L. and Surgailis, D. (1990). *A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittles estimate.* Probab. Theory Related Fields, **vol.** 86, pp. 87-104.

[29] Hall, P. (1997). *On defining and measuring long-range dependence.* Fields Institute Communications, **vol.** 11, pp. 153-160.

[30] Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

[31] Heyde, C.C., Yang, Y. (1997). *On Defining Long-Range Dependence.* Journal of Applied Probability, **vol.** 34, No. 4, pp. 939-944.

[32] Hannan, E.J. (1976). *The asymptotic distribution of serial covariances.* , **vol.** 4, pp. 396-399.

[33] Horváth, L. and Kokoszka, P. (2008). *Sample autocovariances of long-memory time series.* Bernoulli, **vol.** 14, pp. 405-418.

[34] Hurst, H. (1951). *Long-term storage capacity of reservoirs.* Transactions of the American Society of Civil Engineers, **vol.** 116, pp. 770-808.

[35] Hurst, H. (1955). *Methods of using long-term storage in reservoirs.* Proceedings of the Institution of Civil Engineers, Part I, pp. 519-577.

[36] Ibragimov, I.A. and Linnik, Y.V. (1971). *Independent and stationary sequences of random variables.* Wolters-Nordhoff,Groningen.

[37] Karmeshu, D. and Krishnamachari, A. (2004). *Sequence variability and long-range dependence in DNA: An information theoretic perspective.* in Neural Information Processing, pp. 13541361, Berlin: Springer. 3316 of Lecture Notes in Computer Science.

[38] Kolmogorov, A. (1940). *Wienersche Spiralen und einige andere interessante kurven in Hilbertschen raum.* Computes Rendus (Doklady) Academic Sciences USSR (N.S.), **vol.** 26, pp. 115-118.

[39] Kouritzin, M.A. (1996). *On the interrelation of almost sure invariance principles for certain stochastic adaptive algorithms and for partial sums of random variables.* Journal of Theoretical Probability, **vol.** 9, No. 4, pp. 811-840.

[40] Lin, Z. and Lu, C. (1996). *Limit Theory for Mixing Dependent Random Variables.* Kluwer, Dordrecht.

[41] Ljung, L., Soderstrom, T. (1983). *Theory and Practice of Recursive Identification.* MIT Press.

[42] Louhichi S. (2000). *Convergence rates in the strong law for associated random variables.* Probab. Math. Stat., **vol.** 20 pp. 203-214.

[43] Louhchi, S. and Soulier, P. (2000). *Marcinkiewicz-Zegmond Strong Laws for Infinite Variance Time Series.* Statistical Inference for Stochastic Processes, **vol.** 3, pp. 31-40.

[44] Mandelbrot, B. and Wallis, J. (1968). *Noah, Joseph and operational hydrology.* Water Resources Research, **vol.** 4, pp. 909-918.

[45] Mandelbrot, B. and Van Ness, J. (1968). *Fractional Brownian motions, fractional noises and applications.* SIAM Review, **vol.** 10, pp. 422-437.

[46] Mandelbrot, B. (1963). *New methods in statistical economics.* Journal of Political Economy, **vol.** 71, pp. 421-440.

[47] Markov, A.A. (1906). *Rasprostranenie zakona bol'shih chisel na velichiny.* zavisyaschie drug ot druga, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, **vol.** 15 (94), pp. 135-156.

[48] McCulloch, J.H. (1986). *Simple consistent estimators of stable distribution parameters.* Communications in Statistics Simulations, **vol.** 15, pp. 74-81.

[49] Mittniks, S. and Rachevs, S.T. (1997). *Modeling Financial Assets with Alternative Stable Models.* Wiley, New York.

[50] Mittnik, S., and Rachev, S.T. (1993a). *Modeling asset returns with alternative stable models.* Econometric Reviews, **vol.** 12, pp. 261-330.

[51] Mittnik, S., and Rachev, S.T. (1993b). *Reply to comments on "Modeling asset returns with alternative stable models" and some extensions.* Econometric Reviews, **vol.** 12, pp. 347-389.

[52] Mittnik, S., Paolella, M.S., Rachev, S.T. (2000). *Diagnosing and treating the fat tails in financial returns data.* Journal of Empirical Finance, **vol.** 7, pp. 389-416

[53] Mittnik, S., Paolella, M.S., Rachev, S.T. (2002). *Stationary of stable power-GARCH processes.* Journal of Econometrics, **vol.** 106, pp. 97-107.

[54] Mokkadem, A. (1988). *Mixing Properties of ARMA Processes.* Stochastic Processes and their Applications, **vol.** 29, pp. 309-315.

[55] Painter, S. (1995). *Random fractal models of heterogeneity: the Lévy-stable approach.* Math. Geol., **vol.** 27, pp. 813-830.

[56] Painter, S. (1998). *Long-range dependence in the subsurface: Empirical evidence and simulation methods.* Invited paper at the American Geophysical Union, Fall Meeting.

[57] Poisson, S.D. (1837). *Récherches sur la probabilité des jugements en matiére criminelle et en matiére civile.*

[58] Racheva, B. and Samorodnitsky, G. (2003). *Long range dependence in heavy tailed stochastic processes, in Handbook of Heavy Tailed Distributions in Finance.* (edited by Rachev, S. (2003)). Elsevier: Amsterdam.

[59] Phillips, P.C.B. and Solo, V. (1992). *Asymptotics for linear Processes.* The Annals of Statistics, **vol.** 20, pp. 971-1001.

[60] Philipp, W. (1986). *.Invariance principles for independent and weakly dependent random variables. In Dependence in Probability and Statistics: A Survey of Recent Results (E. Eberlein and M. S. Taqqu, eds.)* Birkhäuser, Boston.

[61] Philipp, W. and Stout, W. (1975). *Almost sure invariance principles for partial sums of weakly dependent random variables.* Mem. Amer. Math. Soc. 2.

[62] Rachev, S., Menn, C. and Fabozzi, F. (2005). *Fat-Tailed and Skewed Asset Return Distributions.* John Wiley & Sons. Inc: Hoboken.,**vol.** 12, pp. 261-330.

[63] Rio, E. (1995). *A maximal inequality and dependent Marcinkiewicz-Zygmund strong laws.* Ann. Probab., **vol.** 2(23), pp. 918-937.

[64] Robinson, P. (ed.) (2003). *Time Series with Long Memory. Advanced Texts in Econometrics.* Oxford University Press.

[65] Rosenblatt, M. (1984). *Stochastic processes with short-range and long-range dependence.* in Statistics: An Appraisal, (H. David and H. David, eds.), pp. 509520, Iowa State University Press.

[66] Samorodnitsky, G. (2006). *Long memory and self-similar processes.* Annales de la Faculté des Sciences de Toulouse, **vol.** 15, pp. 107-123.

[67] Samorodnitsky, G. (2002). *Long Range Dependence, Heavy Tails and Rare Events.* MaPhySto, Centre for Mathematical Physics and Stochastics, Aarhus. Lecture Notes.

[68] Saridis, G.N. (1974). *Stochastic Approximation Methods for Identification and Control A Survey.* IEEE-AC, **vol.** 19, No 6.

[69] Shao, Q.M. (1993). *Almost sure invariance principles for mixing sequences of random variables.* Stochastic Process. Appl., **vol.** 48, pp. 319-334.

[70] Stout, W. F. (1974). *Almost Sure Convergence.* Academic Press, New York.

[71] Tadić, V.B. (2004). *On the Almost Sure Rate of Convergence of Linear Stochastic Approximation Algorithms.* IEEE Trans. Inform. Theory, **vol**. 50, No. 2, pp. 401-409.

[72] Taqqu, M. (1986). *A bibliographical guide to self-similar processes and long-range dependence.* in Dependence in Probability and Statistics, (E. Eberlein and M. Taqqu, eds.), pp. 137162, Boston: Birkhäuser.

[73] Vaiciulis, M. (2003). *Convergence of sums of Appell polynomials with infinite variance.* Lithuanian Math. J., **vol.** 43, pp. 80-98.

[74] Varotsos, C. and Kirk-Davidoff, D. (2006). *Long-memory processes in global ozone and temperature variations at the region* 60^0 *S*-60^0 *N.* Atmospheric Chemistry and Physics, **vol.** 6, pp. 4093-4100.

[75] de Vries, C.G. (1991). *On the relation between GARCH and stable processes.* Journal of Econometrics, **vol.** 48, pp. 313-324.

[76] WITHERS, C.S. (1981). *Conditions for linear processes to be strong-mixing.* Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete [Became: @J(ProbTher)], **vol.** 57, pp. 477-480.

[77] WIDROW, B., GLOVER, J.R., MCCOOL, J., KAUNITZ, J., WILLIAMS, C. ., HEARN, R.H., ZEIDLER, J.R., DONG, E., GOODLIN, R.C. (1975). *Adaptive Noise Cancelling : Principles and Applications.* Proc. IEEE, **vol.** 63, No 12.

[78] WU, W.B. (2007). *Strong Invariance Principles for Dependent Random Variables.* The Annals of Probability, **vol.** 35, pp. 2294-2320.

[79] YAGLOM,A. (1955). *Correlation theory of processes with stationary random increments of order n.* Matematica Sbornik, **vol.** 37, pp. 141-196. (English translation in American Mathematical Society of Translations Series 2, **vol.** 8, pp. 87-141, 1958).

Chapter 2

Marcinkiewicz Law of Large Numbers for Outer-products of Heavy-tailed, Long-range Dependent Data [*]

2.1 Introduction

Let $D_k = X_k \overline{X}_k^T$ be random matrices with $\{X_k\}$, $\{\overline{X}_k\}$ being $\mathbb{R}^d$-valued (possibly two-sided, multivariate) linear processes

$$X_k = \sum_{l=-\infty}^{\infty} C_{k-l}\Xi_l, \quad \overline{X}_k = \sum_{l=-\infty}^{\infty} \overline{C}_{k-l}\overline{\Xi}_l. \tag{2.1}$$

[*] A version of this chapter has been accepted for publication. Kouritzin, M.A. and Sadeghi, S. (2015). *Marcinkiewicz Law of Large Numbers for Outer-products of Heavy-tailed, Long-range-Dependence Data.* Advances in Applied Probability Journal, in press.

defined on some probability space (Ω, F, P).

$$\left\{ \left(\Xi_l = (\xi_l^{(1)}, ..., \xi_l^{(m)}), \overline{\Xi}_l = (\overline{\xi}_l^{(1)}, ..., \overline{\xi}_l^{(m)}) \right),\ l \in \mathbb{Z} \right\}$$

are i.i.d. zero-mean random $\mathbb{R}^{m+m}$-vectors (innovations) such that $E[|\Xi_1|^2] < \infty$, $E[|\overline{\Xi}_1|^2] < \infty$ and $(C_l)_{l\in\mathbb{Z}}$, $(\overline{C}_l)_{l\in\mathbb{Z}}$ are $\mathbb{R}^{d\times m}$-matrix sequences satisfying $\sup\limits_{l\in\mathbb{Z}} |l|^{\sigma} \|C_l\| < \infty$, $\sup\limits_{l\in\mathbb{Z}} |l|^{\overline{\sigma}} \|\overline{C}_l\| < \infty$ for some $(\sigma, \overline{\sigma}) \in \left(\frac{1}{2}, 1\right]$. Hence, $\{D_k\}$ can have heavy tails as well as long-range dependence.

Linear process models are heavily used in finance, engineering, econometrics, and statistics. In fact, classical time-series theory mainly involves the statistical analysis of stationary linear processes. Current applications in network theory and financial mathematics leads us to study time series models where $\{D_k\}$ can have heavy tails and long memory. Heavy-tailed data exhibits frequent extremes and infinite variance, while positively-correlated long memory data displays great serial momentum or inertia. Heavy-tailed data with long-range dependence has been observed in a plethora of empirical data set over the last fifty years and so. For instance, Mandelbrot [12] observed that long memory time series often were heavy-tailed and self-similar.

The possible rates of the convergence is affected by both long-range dependence and heavy-tails. There are two broad types of dependence for linear processes. If the coefficients (C_l) are absolutely summable and innovations have second moments, then the covariances of X_k are summable and we say that $\{X_k\}$ is short-range dependence (SRD). On the contrary, we generically say that $\{X_k\}$ is long-range dependence (LRD) if its covariances are not absolutely summable. Practically, by choosing appropriate coefficients, matrix sequence (C_l) can decay slowly enough (as $|l| \to \infty$) such that $\{X_k\}$ shows

LRD. We consider $\{D_k\}$ to have LRD too in this $\{C_l\}$ non-summable case even though the second moments for D_k may not exist. There are also two general kinds of randomness. If each D_k fails to have a second moment, then we say it has heavy-tailed (HT) and is otherwise light-tailed (LT). In our setting, D_k will either have HT or LT depending upon the moments of and dependence between Ξ_1 and $\overline{\Xi}_1$.

There few general Marcinkiewicz Strong Law of Large Numbers (MSLLN) results for partial sums of X_k under both heavy-tailed and the long-range dependence and the MSLLN for partial sums of nonlinear functions of X_k is almost untouched. Our purpose here is to establish a method and a structure under which certain MSLLN for heavy-tailed and the long-range-dependent phenomena can be handled properly. Technically, our goal is to prove:

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (D_k - D) = 0 \quad \text{a.s.} \quad \text{for} \quad p < \frac{1}{2-\sigma-\overline{\sigma}} \wedge \alpha \wedge 2,$$

when $\max\limits_{1\le i,j\le m} \sup\limits_{t\ge 0} t^{\alpha} P(|\xi_1^{(i)}\overline{\xi}_1^{(j)}| > t) < \infty$ for some $\alpha > 1$ and $\sup\limits_{l\in\mathbb{Z}} |l|^{\sigma}\|C_l\| < \infty$, $\sup\limits_{l\in\mathbb{Z}} |l|^{\overline{\sigma}}\|\overline{C}_l\| < \infty$ for some $(\sigma, \overline{\sigma}) \in \left(\frac{1}{2}, 1\right]$. This format of $\{D_k\}$ is critical for our result since it allows the LRD and HT conditions decouple and convergence rate be determined by the worst of the HT requirement $p < (\alpha \wedge 2)$ and the LRD condition $p < \frac{1}{2-\sigma-\overline{\sigma}}$, but not the combination. A bifurcation happens. Consider the summation, $D_k = \sum\limits_{l,m=-\infty}^{\infty} C_{k-l}\Xi_l \overline{C}_{k-m}\overline{\Xi}_m$, broken into off-diagonal and diagonal terms. Due to the independence of $(\Xi_l, \overline{\Xi}_l)$ from $(\Xi_m, \overline{\Xi}_m)$, the off-diagonal sum $\sum\limits_{l\neq m} C_{k-l}\overline{C}_{k-m}\Xi_l\overline{\Xi}_m$ does not have heavy tails (when $\alpha > 1$). Conversely, since $\sigma+\overline{\sigma} > 1$ the diagonal sum $\sum\limits_{l=-\infty}^{\infty} C_{k-l}\overline{C}_{k-l}\Xi_l\overline{\Xi}_l$

does not experience long-range dependence. In addition, the rate of convergence depends on the worst of $(\alpha \wedge 2)$ and $\frac{1}{2-\sigma-\overline{\sigma}}$, so whenever we are in the LRD dominant case, $(\alpha > \frac{1}{2-\sigma-\overline{\sigma}})$, the off-diagonal terms dictate the rate of convergence by the LRD effect $(p < \frac{1}{2-\sigma-\overline{\sigma}})$ and in the HT dominant case, $(\alpha < \frac{1}{2-\sigma-\overline{\sigma}})$, the diagonal terms dictate the rate of convergence by HT effect $(p < \alpha)$. The bifurcation point is when $\alpha = \frac{1}{2-\sigma-\overline{\sigma}}$ and $\alpha < 2$.

2.2 Background

We give a review of some existing literature on MSLLN or weak convergence for partial sums, sample covariance and non-linear function of partial sums with heavy-tailed and/or long-range dependence. Many existing results were only established in the scalar case. For ease of assimilation we use $\{x_k\}$, (c_l), $\{d_k\}$ and $\{\xi_k\}$ to denote these scalar versions of $\{X_k\}$, (C_l), $\{D_k\}$ and $\{\Xi_k\}$ and $\{x_{k+h}\}$ for $\{\overline{X}_k\}$ when it is a shifted version of $\{x_k\}$.

2.2.1 Partial Sums

There are are only a few publication, like Louhchi and Soulier [11], that considered the combination of these LRD and HT phenomena. They stated the following result for linear symmetric α-stable (SαS) processes.

Theorem 2.1 *Let $\{\xi_j\}_{j\in\mathbb{Z}}$ be i.i.d. sequence of SαS random variables with $1 < \alpha < 2$ and $\{c_j\}_{j\in\mathbb{Z}}$ be a bounded collection such that $\sum\limits_{j\in\mathbb{Z}} |c_j|^s < \infty$ for some $s \in [1, \alpha)$. Set $x_k = \sum\limits_{j\in\mathbb{Z}} c_{k-j}\xi_j$. Then, for $p \in (1, 2)$ satisfying $\frac{1}{p} > 1 - \frac{1}{s} + \frac{1}{\alpha}$*

$$\frac{1}{n^{\frac{1}{p}}} \sum_{i=1}^{n} x_i \to 0 \ a.s.$$

The condition $s < \alpha$ ensures $\sum\limits_{j \in \mathbb{Z}} |c_j|^\alpha < \infty$ and thereby convergence of $\sum\limits_{j \in \mathbb{Z}} c_{k-j}\xi_j$. Moreover, $\{x_k\}$ not only exhibits heavy tails but also long-range dependence if, for example, $c_j = |j|^{-\sigma}$ for $j \neq 0$ and some $\sigma \in \left(\frac{1}{2}, 1\right)$. Notice there is interactions between the heavy tail condition and the long-range dependent condition. In particular for a given p, heavier tails (α becomes smaller) implies that you can not have as long-range dependence (s must become smaller) and vice-versa. Moreover, this result does not apply in our outer product setting due to the fact that x_k's are linear processes with SαS innovations and so x_k cannot be decomposed to product of two variables even in the scalar case.

2.2.2 Non-linear function of partial sums

The limit behavior of suitably normalized partial sums of stationary random variables with either LRD or HT has been studied by many authors. Applications can be found in geophysics, economics, hydrology and statistics. For instance, in contexts like Whittle approximation, the asymptotic behavior of quadratic forms of stationary sequences have an important role. In addition, the efficacy of "R/S-statistic" theory that was introduced for estimating the long-run, non-periodic statistical dependence of time series by Hurst and developed by Mandelbrot [13], can be confirmed by convergence of these limit functions.

Many results deal with the existence and description of limit distributions of sums

$$S_{n,h}(t) = \sum_{k=1}^{[nt]} (h(x_k) - E(h(x_k))), \quad t \geq 0, \tag{2.2}$$

where h is a (nonlinear) function. The limit behavior for a Gaussian LRD process $\{x_k\}$, firstly was studied by Rosenblatt [14]. Afterward, Dobrushin and Major [3] explained it in more general form. Then, Taqqu [18] showed that the limit in distribution of particular normalized sums $S_{n,h}(t)$ is determined by the Hermite rank $m^* \in \{1, 2, ...\}$ of $h(x)$, which is the index of the first nonzero coefficient in the Hermite expansion. Nonetheless, the behavior of nonlinear non-Gaussian LRD processes is less known. One of the most studied non-Gaussian LRD processes is the one-sided linear (moving average) process,

$$x_k = \sum_{j=0}^{\infty} c_j \xi_{k-j}, \tag{2.3}$$

in which, innovations $\xi_k, k \in \mathbb{Z}$, are i.i.d., have zero mean with finite variance, and coefficients c_j satisfy:

$$c_j \sim c_\sigma j^{-\sigma}, \quad j \geq 1 \tag{2.4}$$

for some constant $c_\sigma \neq 0$, $c_0 = 1$ and $\sigma \in (\frac{1}{2}, 1)$.

Surgailis [16] considered the limit behavior of partial sum processes $S_{n,h}(t)$ of polynomial h of linear process $\{x_k\}_{k\in\mathbb{Z}}$. Later, Giraitis and Surgailis [4][5], Avram and Taqqu [1] noticed that the only difference between this case and Gaussian case is that the Hermite rank m^* of $h(x)$ has to be replaced by the Appell rank m.

Vaiciulis [19] investigated distributional convergence for normalized partial sums of Appell polynomials $A_m(x_k)$ of linear processes x_k having both long-memory and heavy-tails in the sense $EA_m^2(x_k) = \infty$. In particular, he assumed x_k had the form (2.3) with innovations $\{\xi_k^m\}$ belonging to the domain of attraction of an α-stable law with $1 < \alpha < 2$ and c_j following (2.4). The limit was: **i)** an α-stable Levy process, **ii)** an m^{th} order Hermite process, or **iii)** the sum of two mutually independent α-stable Levy and m^{th} order Hermite processes, depending on the value of α, m and σ where $\sigma \in (\frac{1}{2}, 1)$.

Later, Surgailis [17] considered the bounded, infinitely differentiable h case where $\{x_k\}$ was LRD and had innovations with probability tail decay of $x^{-2\alpha}$ for $1 < \alpha < 2$. Suppose x_k satisfies (2.3) and (2.4). Then, he showed three different limiting behaviors corresponding to three different LRD-HT setting: $n^{1-(2\sigma-1)m^*/2} S_{n,h}(t)$, $n^{\frac{1}{2\alpha\sigma}} S_{n,h}(t)$ or $n^{\frac{1}{2}} S_{n,h}(t)$ converge in distribution to respectively a Hermite process of order m^*, a $2\alpha\sigma$-stable Levy process or a Brownian motion, all at time t, for certain range of α and σ.

2.2.3 Sample Covariances

Auto-covariance functions play a substantial role in time series analysis and have diverse applications in inference problems, including hypothesis testing and parameter estimation. The natural estimator of auto-covariance is sample covariance. Hence, the convergence properties of the sample covariance is of great interest (see, e.g. [2], [6] and [10]). In the case of LRD and HT, it is an area of active research.

Davis and Resnick [2] studied the distributional convergence of sample autocovariances for two-sided linear processes with innovations that were i.i.d. and had regularly varying tail probabilities of index $\alpha > 0$.

$$\begin{aligned}&P(|\xi_k| > x) = x^{-2\alpha} L(x),\\ &\frac{P(\xi_k > x)}{P(|\xi_k| > x)} \to p \quad \text{and} \quad \frac{P(\xi_k < -x)}{P(|\xi_k| > x)} \to q, \quad \text{as } x \to \infty, \end{aligned} \tag{2.5}$$

where $L(.)$ is slowly varying at infinity $\left(\text{so} \lim_{j\to\infty} \dfrac{L(aj)}{L(j)} = 1\right)$ and $0 \le p \le 1$, $q = 1 - p$. They considered the case where the innovations had finite variance (ι) but infinite fourth moment, i.e. $1 < \alpha < 2$ with absolutely summable coefficients c_j with form of (2.4).

Note: We choose to scale our constants, here and in the sequel, so that $\alpha < 2$ always mean HT of the object of interest, which is $x_k x_{k+h}$ or more generally, $X_k \overline{X}_k$.

In case of infinite fourth moment for $\{\xi_k\}_{k\in\mathbb{Z}}$, the asymptotic distribution of normalized sample autocovariances of long-memory processes was studied by Horváth and Kokoszka [6]. Suppose we observe the realization $x_1, x_2, ..., x_{n+v}$, $n > 1, v \ge 0$, the sample autocovariances and population autocovariances are defined as

$$\hat{\gamma}_h^{(n)} = \frac{1}{n}\sum_{k=1}^{n} x_k x_{k+h}, \quad \text{and} \quad \gamma_h = E[x_0 x_h] = \iota \sum_{j=0}^{\infty} c_j c_{j+h},\ h = 0, 1, ..., v, \tag{2.6}$$

respectively. Horváth and Kokoszka [6, Theorem 3.1] studied the asymptotic distribution $[\hat{\gamma}_h^{(n)} - \gamma_h]$, $h = 0, 1, ..., v$ for linear process of form (2.3) with coefficients and innovations satisfying (2.4) and (2.5) and a norming constant $a_n = \inf\{x : P(|\xi_1| > x) \le n^{-1}\}$ (roughly of order $n^{\frac{1}{2\alpha}}$) satisfying

$$\lim_{n\to\infty} nP[|\xi_k| > a_n x] = x^{-2\alpha},\ x > 0. \tag{2.7}$$

We quote this result in our notations as the following theorem.

Theorem 2.2 *Suppose, conditions (2.3), (2.4), (2.5) and (2.7) hold.*

(a) *If* $1 - \frac{1}{2\alpha} < \sigma < 1$ *and* $1 < \alpha < 2$, *then*

$$na_n^{-2}[\hat{\gamma}_h^{(n)} - \gamma_h] \xrightarrow{d} \left(S - \tfrac{\alpha}{\alpha-1}\right)\left[\sum_{j=0}^{\infty} c_j c_{j+h}\right], \quad h = 0, 1, ..., H.$$

where S is an α-stable random variable. (For the above to hold for $\sigma = 3/4$, we must additionally assume that $a_n^{-4} n \ln n \to 0$.)

(b) *If $\frac{1}{2} < \sigma < 1 - \frac{1}{2\alpha}$ and $1 < \alpha < 2$, then*

$$n^{2\sigma-1}[\hat{\gamma}_h^{(n)} - \gamma_h] \xrightarrow{d} \iota c_\sigma^2 \left[U_\sigma(1)\right], \quad h = 0.1, ..., H.$$

where U_σ is a Rosenblatt process.

The Rosenblatt process is often defined by the iterated stochastic integral:

$$U_\sigma(t) = 2\int_{w_1<w_2<t}\left[\int_0^t (\tau - w_1)_+^{-\sigma}(\tau - w_2)_+^{-\sigma} d\tau\right] W(dw_1)W(dw_2),$$

in which $W(.)$ is the standard Wiener process on the real line.

This theorem works for one-sided linear processes with a regularly varying tail condition and gives us weak convergence.

Notice that in Theorem 2.2, case **(a)** represents the HT dominant, ($\alpha < \frac{1}{2-2\sigma}$), so the diagonal terms dictate convergence to an α-stable distribution. However, case **(b)** represents the LRD dominant, ($\alpha > \frac{1}{2-2\sigma}$), hence off-diagonal terms take over and we get convergence to Rosenblatt process.

2.3 Main Results

Our first result is in the scalar case. Later, we will extract the vector-valued result as a second theorem. All proofs are delayed until the next section after our applications.

Theorem 2.3 *Let* $\left\{(\xi_l, \overline{\xi}_l)\right\}_{l\in\mathbb{Z}}$ *be i.i.d. zero-mean random variables such that* $E[\xi_1^2] < \infty$, $E[\overline{\xi}_1^2] < \infty$ *and* $\sup\limits_{t\geq 0} t^\alpha P(|\xi_1\overline{\xi}_1| > t) < \infty$ *for some* $\alpha > 1$. *Moreover, suppose* $(c_l)_{l\in\mathbb{Z}}, (\overline{c}_l)_{l\in\mathbb{Z}}$ *satisfy* $\sup\limits_{l\in\mathbb{Z}} |l|^\sigma |c_l| < \infty$, $\sup\limits_{l\in\mathbb{Z}} |l|^{\overline{\sigma}} |\overline{c}_l| < \infty$ *for some* $\sigma, \overline{\sigma} \in \left(\frac{1}{2}, 1\right]$, $d_k = \sum\limits_{l,m=-\infty}^{\infty} c_{k-l}\overline{c}_{k-m}\xi_l\overline{\xi}_m$ *and* $d = E[\xi_1\overline{\xi}_1] \sum\limits_{l=-\infty}^{\infty} c_l\overline{c}_l$. *Then, for p satisfying* $p < \frac{1}{2-\sigma-\overline{\sigma}} \wedge \alpha \wedge 2$

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (d_k - d) = 0 \ a.s.$$

Remark 2.1 *The tail probability bound ensures that* $E[|\xi_1\overline{\xi}_1|^r] < \infty$ *for any* $r \in (1, (\alpha \wedge 2))$ *and* $E[d_1]$ *exists but it is possible that* $E[d_1^2] = \infty$ *so, we are handling heavy tails for* $\{d_k\}$. *On the other hand,* $E[|\xi_1\overline{\xi}_1|^\alpha] < \infty$ *implies our tail condition by Markov's inequality.* σ, $\overline{\sigma}$ *bound the amount of long-range dependence in* $x_k = \sum\limits_{l=-\infty}^{\infty} c_{k-l}\xi_l$, $\overline{x}_k = \sum\limits_{l=-\infty}^{\infty} \overline{c}_{k-l}\overline{\xi}_l$. *If* σ *can be taken larger than* 1, *then* $\sum\limits_{k=1}^{\infty} E[x_0 x_k] < \infty$ *and there is no long-range dependence in* $\{x_k\}$. $\sigma > \frac{1}{2}$ *with* $E[\xi_1^2] < \infty$ *ensures that* $\sum\limits_{l=-\infty}^{\infty} c_{k-l}\xi_l$ *converges a.s.*

Remark 2.2 *Notice that the constraints to handle long-range dependence,* $p < \frac{1}{2-\sigma-\overline{\sigma}}$, *and to handle the heavy tails,* $p < (\alpha \wedge 2)$, decouple. *This decoupling appears to be due to the structure of* d_k. *Due to the independence of* $(\xi_l, \overline{\xi}_l)$ *from* $(\xi_m, \overline{\xi}_m)$, *the off-diagonal sum* $\sum\limits_{l\neq m} c_{k-l}\overline{c}_{k-m}\xi_l\overline{\xi}_m$ *does not have heavy tails. Conversely, since* $\sigma + \overline{\sigma} > 1$ *the diagonal sum* $\sum\limits_{l=-\infty}^{\infty} c_{k-l}\overline{c}_{k-l}\xi_l\overline{\xi}_l$ *does not experience long-range dependence.*

Corollary 2.1 *Assume* $\ell(\cdot), L(\cdot)$ *be slowly varying functions,* $\left\{(\xi_l, \overline{\xi}_l)\right\}_{l\in\mathbb{Z}}$ *be i.i.d. zero-mean random variables such that* $E[\xi_1^2] < \infty$, $E[\overline{\xi}_1^2] < \infty$ *and* $\sup\limits_{t\geq 0} t^\alpha P(|\xi_1\overline{\xi}_1| > t) < \infty$ *for some* $\alpha > 1$. *Also, suppose* $(c_l)_{l\in\mathbb{Z}}, (\overline{c}_l)_{l\in\mathbb{Z}}$ *sat-*

isfy $\sup_{l\in\mathbb{Z}} \ell(l)|l|^{\sigma}|c_l| < \infty$, $\sup_{l\in\mathbb{Z}} \ell(l)|l|^{\overline{\sigma}}|\overline{c}_l| < \infty$, *for some* $\sigma, \overline{\sigma} \in \left(\frac{1}{2}, 1\right]$, $d_k = \sum\limits_{l,m=-\infty}^{\infty} c_{k-l}\overline{c}_{k-m}\xi_l\overline{\xi}_m$ *and* $d = E[\xi_1\overline{\xi}_1] \sum\limits_{l=-\infty}^{\infty} c_l\overline{c}_l$. *Then, for* p *satisfying* $p < \frac{1}{2-\sigma-\overline{\sigma}} \wedge \alpha \wedge 2$, $\lim\limits_{n\to\infty} \frac{L(n)}{n^{\frac{1}{p}}} \sum\limits_{k=1}^{n} (d_k - d) = 0$ *a.s.*

Proof. Inasmuch as the two generalizations to allow slowly varying functions are similar, we just illustrate one. Let $p' \in (p, \frac{1}{2-\sigma-\overline{\sigma}} \wedge \alpha \wedge 2)$ so $\lim\limits_{n\to\infty} \frac{1}{n^{\frac{1}{p'}}} \sum\limits_{k=1}^{n} (d_k - d) = 0$ a.s. by Theorem 2.3. By the representation theorem for the slowly varying function $L(n) = \exp\left(\eta(n) + \int_B^n \frac{\theta(t)}{t} dt\right)$ for some $B > 0$ and bounded measurable η, θ with $\lim\limits_{n\to\infty} \eta(n)$ existing and $\lim\limits_{t\to\infty} \theta(t) = 0$. Hence,

$$\lim_{n\to\infty} \frac{L(n)}{n^{\frac{1}{p}}} n^{\frac{1}{p'}} = \lim_{n\to\infty} \exp\left(\eta(n) + \int_B^n \frac{\theta(t)}{t} dt + \left(\frac{1}{p'} - \frac{1}{p}\right) \int_1^n \frac{1}{t} dt\right) = 0. \square$$

We will give a simple example to verify conditions in Theorem 2.3. Recall, a non-negative random variable ξ obeys a power law with parameters $\beta > 1$ and $x_{\min} > 0$, written $\xi \sim PL(x_{min}, \beta)$, if it has density

$$f(x) = \frac{\beta - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\beta} \quad \forall\ x \geq x_{min}$$

so $E|\xi|^r = \begin{cases} x_{min}^r\left(\frac{\beta-1}{\beta-1-r}\right) & r < \beta - 1 \\ \infty & r \geq \beta - 1 \end{cases}$.

It has a folded t distribution with parameter $\beta > 1$, written $\xi \sim Ft(\beta)$, if it has density

$$f(x) = \frac{2\Gamma(\frac{\beta}{2})}{\Gamma(\frac{\beta-1}{2})\sqrt{(\beta-1)\pi}} \left(1 + \frac{x^2}{(\beta-1)}\right)^{-\frac{\beta}{2}} \quad \forall\ x > 0$$

so $E(|\xi|^r)$ exists if and only if $r < \beta - 1$.

Example 2.1 *Suppose $p, q, \alpha, \beta, \overline{\beta} > 1$ are such that $\frac{1}{p} + \frac{1}{q} = 1$, $\beta > p\alpha + 1$, $\overline{\beta} > q\alpha + 1$ and $p\alpha, q\alpha \geq 2$. If ξ_1 and $\overline{\xi_1}$ have power law distribution, lets say $\xi_1 \sim Pl(x_{\min}, \beta)$, $\overline{\xi}_1 \sim Pl(\overline{x}_{\min}, \overline{\beta})$ for some $x_{\min}, \overline{x}_{\min} > 0$, then $E[\xi_1^2]$, $E[\overline{\xi}_1^2] < \infty$ and $\sup\limits_{t \geq 0} t^\alpha P(|\xi_1 \overline{\xi}_1| > t) < \infty$. If $\xi_1 \sim Ft(\beta)$, $\overline{\xi}_1 \sim Ft(\overline{\beta})$, then $E[\xi_1^2]$, $E[\overline{\xi}_1^2] < \infty$ and $\sup\limits_{t \geq 0} t^\alpha P(|\xi_1 \overline{\xi}_1| > t) < \infty$. Either way, the Theorem 2.3 applies with properly chosen $(c_l, \overline{c}_l)$.*

We now consider the case where X_k and $\overline{X}_k$ are (multivariate) linear processes.

Theorem 2.4 *Let $\{\Xi_l\}$ and $\{\overline{\Xi}_l\}$ be i.i.d. zero-mean random $\mathbb{R}^m$-vectors such that $\Xi_l = \left(\xi_l^{(1)}, ..., \xi_l^{(m)}\right)$, $\overline{\Xi}_l = \left(\overline{\xi}_l^{(1)}, ..., \overline{\xi}_l^{(m)}\right)$, $\max\limits_{1 \leq i,j \leq m} \sup\limits_{t \geq 0} t^\alpha P(|\xi_1^{(i)} \overline{\xi}_1^{(j)}| > t) < \infty$ for some $\alpha > 1$, $E[|\Xi_1|^2] < \infty$ and $E[|\overline{\Xi}_1|^2] < \infty$. Moreover, suppose matrix sequences $(C_l)_{l \in \mathbb{Z}}, (\overline{C}_l)_{l \in \mathbb{Z}} \in \mathbb{R}^{d \times m}$ satisfy*

$$\sup_{l \in \mathbb{Z}} |l|^\sigma \|C_l\| < \infty, \sup_{l \in \mathbb{Z}} |l|^{\overline{\sigma}} \|\overline{C}_l\| < \infty \quad \textit{for some} \quad (\sigma, \overline{\sigma}) \in \left(\frac{1}{2}, 1\right],$$

X_k, $\overline{X}_k$ take form of (2.1), $D_k = X_k \overline{X}_k^T$ and $D = E[X_1 \overline{X}_1^T]$. Then, for p satisfying $p < \frac{1}{2 - \sigma - \overline{\sigma}} \wedge \alpha \wedge 2$

$$\lim_{n \to \infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (D_k - D) = 0 \qquad a.s.$$

This theorem follows by linearity of limits and Theorem 2.3.

2.3.1 Application: Stochastic Approximation

Stochastic approximation (SA) is often used in optimization problems for linear models. Hence, the convergence properties of SA algorithms driven by linear models is of utmost interest (see, e.g. [7], [8] and [9]). For illustration, we assume $\{z_k, k = 1, 2, ..\}$ and $\{y_k, k = 2, 3, ...\}$ are respectively $\mathbb{R}^d-$ and $\mathbb{R}-$valued stochastic processes, defined on some probability space (Ω, F, P), that satisfy $y_{k+1} = z_k^T h + \epsilon_k$, $\forall k = 1, 2, \ldots$, where h is an unknown d-dimensional parameter or weight vector of interest and $\{\epsilon_k\}$ is a noise sequence. We want to estimate the parameter vector h through the stochastic approximation algorithm:

$$h_{k+1} = h_k + \mu_k(b_k - A_k h_k), \tag{2.8}$$

where μ_k is the k^{th} step gain of the form $\mu_k = k^{-\chi}$ for some $\chi \in \left(\frac{1}{2}, 1\right]$, $A_k = z_k z_k^T$ and $b_k = y_{k+1} z_k$.

Kouritzin and Sadeghi [7] studied the convergence and almost sure rates of convergence for the algorithm (2.8) in a general enough setting to handle HT and LRD. Now, we can combine our main result (Theorem 2.4) with [7, Corollary 3] to obtain a powerful rate of convergence result for stochastic approximation.

Theorem 2.5 *Let $\{\Xi_l\}$ be i.i.d. zero-mean random $\mathbb{R}^m$-vectors such that for some $\alpha \in (1, 2)$, $\sup\limits_{t \geq 0} t^\alpha P(|\Xi_1|^2 > t) < \infty$, $(C_l)_{l \in \mathbb{Z}}$ be $\mathbb{R}^{(d+1)\times m}$-matrices such that $\sup\limits_{l \in \mathbb{Z}} |l|^\sigma \|C_l\| < \infty$ for some $\sigma \in \left(\frac{1}{2}, 1\right]$, $(z_k^T, y_{k+1})^T = \sum\limits_{l=-\infty}^{\infty} C_{k-l}\Xi_l$, $A_k = z_k z_k^T$ and $b_k = y_{k+1} z_k$ and $A = E[z_k z_k^T]$ and $b = E[y_{k+1} z_k]$.*

Then, $|h_n - h| = o(n^{-\gamma})$ as $n \to \infty$ a.s. for any $\gamma < \gamma_0^{(\chi)} \doteq (\chi - \frac{1}{\alpha}) \wedge (\chi + 2\sigma - 2)$.

Proof. By Theorem 2.4 when $\frac{1}{p} = \chi - \gamma$, $\overline{X}_k^T = X_k^T = (z_k^T, y_{k+1})$, $\overline{\Xi}_l = \Xi_l$, $\overline{C}_l = C_l$, $\overline{\sigma} = \sigma$, and $D_k = \begin{pmatrix} z_k z_k^T & y_{k+1} z_k \\ y_{k+1} z_k^T & y_{k+1}^2 \end{pmatrix}$,

$$\frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (D_k - D) \to 0 \quad a.s.,$$

where $D = \begin{pmatrix} A & b \\ b^T & E[y_{k+1}^2] \end{pmatrix}$. The first d-rows of $\frac{1}{n^{\chi-\gamma}} \sum\limits_{k=1}^{n} (D_k - D) \to 0$ a.s. then establish the MSLLN

$$\frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (A_k - A) \to 0 \quad \text{and} \quad \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (b_k - b) \to 0 \quad a.s.$$

Now, we apply [7, Corollary 3] to complete the proof. □

Remark 2.3 *Note that $\chi - \gamma$ satisfies the required conditions $\chi - \gamma > 2 - 2\sigma$ and $\chi - \gamma > \frac{1}{\alpha}$ in Theorem 2.4. Theorem 2.5 also appears in [7, Theorem 9].*

2.3.2 Application: Non-linear Function of Linear Processes

As mentioned in Background, Vaiciulis [19] showed the convergence of distributions of the partial sum processes with non-linear $h(x_k)$ in terms of convergence of Appell polynomials $A_m(x_k)$ of a long-memory moving average process $\{x_k\}$ with i.i.d. innovations $\{\xi_k\}$ in the case where the variance $EA_m^2(x_k) = \infty$, and the distribution of ξ_1^m belongs to the domain of attraction of an α-stable law with $1 < \alpha < 2$.

Practically, the simplest examples of functions $h(x)$ with a given Appell rank m are Appell polynomials $h = A_m$ relative to the marginal distribution x_1 of the linear process (2.3). In case $m = 2$ the Appell polynomial is $A_2(x) = x^2 - \mu_2$ where $\mu_2 = Ex^2$. Viaiciulis [19, Theorems 1.1 and 1.2] proved that when $m(2\sigma - 1) < 1, m \geq 2$ and $\sigma \in (\frac{1}{2}, 1)$ the limit distribution of partial sums of m^{th} Appell polynomial is either **(i)** an α-stable Levy process for $2 - 2\sigma < 1 + \frac{2}{m}(\frac{1}{\alpha} - 1)$, or **(ii)** an m^{th} order Hermite process for $2 - 2\sigma > 1 + \frac{2}{m}(\frac{1}{\alpha} - 1)$ or **(iii)** the sum of two mutually independent processes depending on the value of α, m and σ, for $2 - 2\sigma = 1 + \frac{2}{m}(\frac{1}{\alpha} - 1)$.

Taking into account all his conditions (when $t = 1$) and transforming it to our case we write our complementary almost sure rate-of-convergence theorem:

Theorem 2.6 *Suppose A_2 represents the Appell polynomials with rank 2 relative to the marginal distribution x_1 of the linear process $x_k = \sum_{j=0}^{\infty} c_{k-j}\xi_j$, for $p \in [1, \frac{1}{2-2\sigma} \wedge \alpha)$ when*

$$\sup_{t \geq 0} t^\alpha P(\xi_1^2 > t) < \infty, \ \sup_{l \in \mathbb{Z}} |l|^\sigma |c_l| < \infty, \ \textit{for some } \alpha \in (1,2), \sigma \in \left(\frac{1}{2}, 1\right]. \tag{2.9}$$

Then, $\lim_{n \to \infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} A_2(x_k) = 0$ a.s.

One might wonder if we have obtained the best possible MSLLN. Indeed, we have. For example, Viaiciulis [19] shows convergence in distribution of $\frac{1}{n^{(2-2\sigma)\wedge\frac{1}{\alpha}}} \sum_{k=1}^{n} A_2(x_k)$ for $m = 2$ to different non-trivial limits in cases $(2 - 2\sigma) > \frac{1}{\alpha}$ (LRD dominant) or $(2 - 2\sigma) < \frac{1}{\alpha}$ (HT dominant), respectively. Therefore, $\dfrac{1}{n^{(2-2\sigma)\wedge\frac{1}{\alpha}}} \displaystyle\sum_{k=1}^{n} A_2(x_k)$ cannot converge to zero almost surely. Theorem 2.6 gives MSLLN for Appell polynomials with rank 2 or in other word gives the convergence and almost sure rates of convergence for partial sums of second Appell polynomial when $\frac{1}{p} > (2 - 2\sigma) \vee \frac{1}{\alpha}$. Our result is optimal in polynomial sense and we cannot do better than that in terms of MSLLN.

2.3.3 Application: Autocovariances

As mentioned in the background, autocovariance estimation under HT and LRD conditions is an active area of research. We will handle the asymptotic behavior of sample covariance function for processes with LRD, innovations of infinite 4^{th} moment and finite variance ι. If we define the sample aurtocovariance and population autocovariance functions by $\hat{\gamma}^{(n)}(h)$ and $\gamma(h)$, as (2.6), we have following almost sure result.

Theorem 2.7 *Assume $\hat{\gamma}^{(n)}(h)$ and $\gamma(h)$, as (2.6) in which $x_k = \sum_{j=0}^{\infty} c_{k-j}\xi_j$ and satisfies (2.9) with $E[\xi_1^2] = \iota$. Then, for p satisfying $p < \frac{1}{2-2\sigma} \wedge \alpha \wedge 2$*

$$n^{1-\frac{1}{p}}[\hat{\gamma}_h^{(n)} - \gamma_h] \to 0 \ a.s. \tag{2.10}$$

Proof. Note that in Theorem 2.3, for case $\overline{\xi}_l = \xi_l$, $E[\xi_1^2] = \iota$, $\overline{c}_l = c_{l+h}$ and $\{c_l = 0, \forall\, l < 0\}$ we have

$$d_k = \sum_{l=-\infty}^{k} \sum_{m=-\infty}^{k+h} c_{k-l}c_{k+h-m}\xi_l\xi_m \quad \text{and} \quad d = \iota \sum_{l=0}^{\infty} c_l c_{l+h}.$$

Hence, $\frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (d_k - d) = \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} \left(\sum_{l=-\infty}^{k} \sum_{m=-\infty}^{k+h} c_{k-l}c_{k+h-m}\xi_l\xi_m - \iota \sum_{l=0}^{\infty} c_l c_{l+h} \right)$.

On the other hand, (2.10) can be written as

$$\begin{aligned} n^{1-\frac{1}{p}}[\hat{\gamma}_h^{(n)} - \gamma_h] &= \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (x_k x_{k+h} - E x_0 x_h) \\ &= \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} \left(\sum_{l=-\infty}^{k} \sum_{m=-\infty}^{k+h} c_{k-l}c_{k+h-m}\xi_l\xi_m - \iota \sum_{l=0}^{\infty} c_l c_{l+h} \right). \end{aligned} \tag{2.11}$$

So, the result follows.□

As we saw, Theorem 2.2 gives the convergence to the following non-trivial limits for $\frac{2\alpha-1}{2\alpha} < \sigma < 1$ and $\frac{1}{2} < \sigma < \frac{2\alpha-1}{2\alpha}$ when $1 < \alpha < 2$,

$$\begin{aligned} &\textbf{(a)} \qquad \frac{1}{a_n^2}\sum_{k=1}^{n}(x_k x_{k+h} - Ex_0x_h) \xrightarrow{d} \left(S - \frac{\alpha}{\alpha-1}\right)\left[\sum_{l=0}^{\infty} c_l c_{l+h}\right], \\ &\textbf{(b)} \qquad \frac{1}{n^{2-2\sigma}}\sum_{k=1}^{n}(x_k x_{k+h} - Ex_0x_h) \xrightarrow{d} \iota c_\sigma^2\left[U_\sigma(1)\right], \qquad (2.12) \end{aligned}$$

respectively, for $h = 0, 1, ..., v$.

It is clear that in the case of HT dominant, $\frac{1}{\alpha} > 2-2\sigma$, we have almost-sure convergence (Theorem 2.7) when $\frac{1}{p} > \frac{1}{\alpha}$. When $\frac{1}{p} = \frac{1}{\alpha}$, we get into the case **(a)** and have convergence to an α-stable distribution. On the other hand, in the LRD dominant case, $\frac{1}{\alpha} < 2-2\sigma$, (from Theorem 2.7) we have almost-sure convergence for $\frac{1}{p} > 2-2\sigma$, yet for $\frac{1}{p} = (2-2\sigma)$ we have convergence to Rosenblatt process by **(b)** .

Hence, Theorem 2.7 shows the *a.s* convergence for difference of sample autocovariance and population autocovariance with HT and LRD. One example can be in the case that $h = 0$. Theorem 2.2 and (2.12) give the convergence in distribution

$$\begin{aligned} \frac{1}{a_n^2}\sum_{k=1}^{n}(x_k^2 - Ex_0^2) &\xrightarrow{d} (S - \frac{\alpha}{\alpha-1})\sum_{l=0}^{\infty} c_l^2 \\ \frac{1}{n^{2-2\sigma}}\sum_{k=1}^{n}(x_k^2 - Ex_0^2) &\xrightarrow{d} \iota c_\sigma^2 U_\sigma(1), \end{aligned}$$

for $\frac{1}{p} = \frac{1}{\alpha}$ and $\frac{1}{p} = 2-2\sigma$, respectively.

While, Theorem 2.7 gives the almost-sure convergence for $\frac{1}{n^{\frac{1}{p}}}\sum_{k=1}^{n}\left(x_k^2 - Ex_0^2\right)$ when $\frac{1}{p} > (2-2\sigma) \vee \frac{1}{\alpha}$.

When we have convergence in distribution to non-trivial limits we can not get almost-sure convergence to 0. However, by Theorem 2.7 we can get arbitrary close to that with polynomial rate and get optimal polynomial almost sure rate of convergence. We can not do better than that in terms of MSLLN.

Remark 2.4 *The power to detect long-range dependence and estimate σ can improve by considering* increasing *(in n) lag autocovariances in lieu of fixed lag. As random variables are moved apart the magnitude of their covariance often decreases at a rate that depends on the long-range dependence coefficient σ. Hence, one can wonder if a MSLLN with a faster rate exists for the increasing lag covariance in our* two-*sided linear process setting. Unfortunately, the answer appears to be no. To explain, we let π_n be the lag (so $\overline{x}_k = x_{k+\pi_n}$) and consider the convergence rate of $\frac{1}{n}\sum_{k=1}^{n}\sum_{l,m} c_{k-l}c_{k+\pi_n-m}\xi_l\xi_m$, with π_n non-decreasing and satisfying $0 \le \pi_n \le n$. Then, after some calculus, one finds that the diagonal terms satisfy $\frac{1}{\pi_n^{1-2\sigma}n^{\frac{1}{p}}}\sum_{k=1}^{n}\sum_{l=-\infty}^{\infty} c_{k-l}c_{k+\pi_n-l}\xi_l^2 \to 0$ a.s. (under mild additional assumptions on $\{c_i\}$), implying the order $\pi_n^{1-2\sigma}$ rate increase over our fixed lag results as expected from the results in Wu et. al. [22]. However, letting $\overset{D}{=}$ denote equal in distribution and taking $m' = m - \pi_n$, we find that*

$$\sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}\sum_{\substack{m=-\infty\\ m\neq l,l+\pi_n}}^{\infty} c_{k-l}c_{k+\pi_n-m}\xi_l\xi_m \overset{D}{=} \sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}\sum_{\substack{m=-\infty\\ m\neq l,l+\pi_n}}^{\infty} c_{k-l}c_{k+\pi_n-m}\xi_l\xi_{m-\pi_n}$$
$$= \sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}\sum_{\substack{m'=-\infty\\ m'\neq l,l-\pi_n}}^{\infty} c_{k-l}c_{k-m'}\xi_l\xi_{m'}. \quad (2.13)$$

The terms on the far left is the off-diagonal terms for the increasing lag case with the terms $m = l + \pi_n$ removed whereas the far right is the off-diagonal *terms for the constant lag case with the terms $m' = l - \pi_n$ removed. Now, our off-diagonal term bounds in the proof to follow appear to be of tight order so a 'faster' MSLLN for the increasing lag case seems unlikely. Moreover, the removed terms would only appear in the large c and mixed terms since $\pi_n \le n$. Hence, the bound of $E\left[(S_n^{(3)})^2\right]$ would remain unchanged and $T(n)$ would have to increase faster to expect a better convergence rate. However, that would negatively affect the bound of $E\left[(S_n^{(2)})^2\right]$, which seems insignificantly changed by the single m term swap in the above equation. Consequently, we believe the increasing-lag autocovariance MSLLN has the same form as the constant lag and the increasing lag autocovariance test has limited advantage over the fixed lag in the two-sided linear process case. Thus, we have not complicated our proofs to include the increasing-lag case.*

2.4 Proofs

In this section proof of theorems will be provided.

2.4.1 A First Light Tail Result

We first give a result that only handles long-range dependence without heavy tails. However, our proof of Theorem 2.3 to follow will show that these two phenomena decouple, so we can easily build upon the Theorem 2.8 to handle both long-range dependence and heavy tails together.

Theorem 2.8 *Let $\left\{(\xi_l, \overline{\xi}_l),\ l \in \mathbb{Z}\right\}$ be i.i.d. zero-mean random variables such* *that* $E[(1+\xi_1^2)(1+\overline{\xi}_1^2)] < \infty$, $(c_l, \overline{c}_l)_{l\in\mathbb{Z}}$ *satisfy*

$$\sup_{l\in\mathbb{Z}} |l|^{\sigma}|c_l| < \infty, \quad \sup_{l\in\mathbb{Z}} |l|^{\overline{\sigma}}|\overline{c}_l| < \infty \quad \textit{for some} \quad \sigma, \overline{\sigma} \in \left(\frac{1}{2}, 1\right],$$

$x_k = \sum\limits_{l=-\infty}^{\infty} c_{k-l}\xi_l$, $\overline{x}_k = \sum\limits_{l=-\infty}^{\infty} \overline{c}_{k-l}\overline{\xi}_l$, $d_k = x_k\overline{x}_k = \sum\limits_{l,m=-\infty}^{\infty} c_{k-l}\overline{c}_{k-m}\xi_l\overline{\xi}_m$ *and* $d = E[\xi_1\overline{\xi}_1]\sum\limits_{l=-\infty}^{\infty} c_{k-l}\overline{c}_{k-l} = E[\xi_1\overline{\xi}_1]\sum\limits_{l=-\infty}^{\infty} c_l\overline{c}_l$. *Then, for* $p < \frac{1}{2-\sigma-\overline{\sigma}}$

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (d_k - d) = 0 \quad a.s.$$

Proof. Insomuch as the proof of the general case only differs cosmetically from the notationally-simpler case where $\overline{\xi}_l = \xi_l$ and $\overline{c}_l = c_l = \begin{cases} 1 & l = 0 \\ |l|^{-\sigma} & l \neq 0 \end{cases}$, we only provide the proof of the latter for which the constraint becomes $p < \frac{1}{2-2\sigma}$. Assume without loss of generality that $\sigma < 1$ and $E[\xi_1^2] = 1$.

Step 1: Divide partial sums into diagonal, large c, small and mixed type terms.

Let $n_r = 2^r$ and $T = T(n) = n^{\nu}$ for $\nu > 0$, $n \in [n_r, n_{r+1})$ and $r \in \mathbb{N}_0$, and define

$$S_n^{(1)} = \sum_{k=1}^{n} \sum_{l=-\infty}^{\infty} c_{k-l}^2 \left(\xi_l^2 - 1\right) \tag{2.14}$$

$$S_n^{(2)} = \sum_{k=1}^{n} \sum_{\substack{l,m=k-T \\ l\neq m}}^{k+T} c_{k-l}c_{k-m}\xi_l\xi_m \tag{2.15}$$

$$S_n^{(3)} = \sum_{k=1}^{n} \sum_{\substack{(l-k)\wedge(m-k)>T \\ l\neq m}} c_{k-l}c_{k-m}\xi_l\xi_m \tag{2.16}$$

$$S_n^{(4)} = \sum_{k=1}^{n} \sum_{m-k>T} \sum_{l=k-T}^{k+T} c_{k-l}c_{k-m}\xi_l\xi_m. \tag{2.17}$$

By breaking $\left\{ \frac{1}{n^{\frac{1}{p}}} \sum\limits_{k=1}^{n} (d_k - d)\,,\ n = 1, 2, ... \right\}$ into pieces and considering those pieces with different (process) distributions, we just need to show that

$$\lim_{n\to\infty} \frac{S_n^{(1)}}{n^{\frac{1}{p}}} = \lim_{n\to\infty} \frac{S_n^{(2)}}{n^{\frac{1}{p}}} = \lim_{n\to\infty} \frac{S_n^{(3)}}{n^{\frac{1}{p}}} = \lim_{n\to\infty} \frac{S_n^{(4)}}{n^{\frac{1}{p}}} = 0 \text{ a.s.},$$

provided $p < \frac{1}{2-2\sigma}$. To handle (the diagonal terms) $S_n^{(1)}$, we let $\zeta_l = \xi_l^2 - 1$, set $K = E[\zeta_1^2]$ and use standard steps.

Step 2: Bound second moment of geometric diagonal partial sums $S_{n_r}^{(1)}$. By symmetry and then integral approximation, we have that

$$\begin{aligned}
& E[(S_{n_r}^{(1)})^2] \\
= \;& \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \sum_{j=1}^{n_r} \sum_{k=1}^{n_r} c_{k-l}^2 c_{j-m}^2 E[\zeta_l \zeta_m] \\
= \;& K \sum_{l=-\infty}^{\infty} \left| \sum_{k=1}^{n_r} c_{k-l}^2 \right|^2 \\
\overset{r}{\ll} \;& \sum_{k=1}^{n_r} \left(1 + 2 \sum_{l=1}^{\infty} l^{-4\sigma} + 2 \sum_{j=k+1}^{n_r} \left(2(j-k)^{-2\sigma} + \sum_{l=-\infty}^{k-1} (k-l)^{-2\sigma} (j-l)^{-2\sigma} \right.\right. \\
+ \;& \left.\left. \sum_{l=k+1}^{j-1} (l-k)^{-2\sigma} (j-l)^{-2\sigma} + \sum_{l=j+1}^{\infty} (l-k)^{-2\sigma} (l-j)^{-2\sigma} \right) \right) \\
\overset{r}{\ll} \;& \sum_{k=1}^{n_r} \left(1 + \sum_{j=k+1}^{n_r} \left((j-k)^{-2\sigma} + (j-k)^{1-4\sigma} \right) \right) \overset{r}{\ll} n_r. \qquad (2.18)
\end{aligned}$$

$$\begin{aligned}
\textbf{Note: } \sum_{l=k+1}^{j-1} \frac{1}{(l-k)^{2\sigma} (j-l)^{2\sigma}} \;&\leq\; 2 \sum_{l=k+1}^{\left\lfloor \frac{j+k}{2} \right\rfloor} \frac{1}{(l-k)^{2\sigma} (j-l)^{2\sigma}} \\
&\overset{j,k}{\ll}\; (j-k)^{-2\sigma} \sum_{l=k+1}^{\left\lfloor \frac{j+k}{2} \right\rfloor} \frac{1}{(l-k)^{2\sigma}} \\
&\overset{j,k}{\ll}\; (j-k)^{(1-4\sigma)}. \qquad (2.19)
\end{aligned}$$

Step 3: Maximal bound for geometric diagonal partial sums.

Following (2.18) we have for $n_r \le n < o < n_{r+1}$

$$\begin{aligned} E[(S_o^{(1)} - S_n^{(1)})^2] &\le K \sum_{l=-\infty}^{\infty} \left| \sum_{k=n+1}^{o} c_{k-l}^2 \right|^2 \\ &\overset{o,n}{\ll} \sum_{k=n+1}^{o} \left(1 + \sum_{j=k+1}^{o} ((j-k)^{-2\sigma} + (j-k)^{1-4\sigma}) \right) \\ &\overset{o,n}{\ll} o - n. \end{aligned} \tag{2.20}$$

Therefore, it follows by Theorem 2.4.1 of Stout [15] with $g(a,n) = Cn$ for some constant $C > 0$ that

$$\begin{aligned} E\left[\max_{n_r \le n < o < n_{r+1}} \left(S_o^{(1)} - S_n^{(1)} \right)^2 \right] &\overset{r}{\ll} \left(\frac{\log(2(n_{r+1} - n_r))}{\log 2} \right)^2 (n_{r+1} - n_r) \\ &\overset{r}{\ll} r^2 n_r. \end{aligned} \tag{2.21}$$

Step 4: Use previous two steps to show *normalized* diagonal sums converge.

Combining (2.18) and (2.21), one has that

$$\sum_{r=0}^{\infty} E\left[\max_{n_r \le n < n_{r+1}} \left(\frac{S_n^{(1)}}{n^{\frac{1}{p}}} \right)^2 \right] \ll \sum_{r=0}^{\infty} r^2 n_r^{1-\frac{2}{p}} < \infty, \tag{2.22}$$

provided $p \in (0,2)$. It follows by Fubini's Theorem and n^{th} term divergence that

$$\lim_{n\to\infty} \frac{S_n^{(1)}}{n^{\frac{1}{p}}} = 0.$$

Step 5: Set up for off-diagonal terms.

Letting

$$\begin{aligned} a^{2,n}_{l,m} &= 2\sum_{k=1}^{n} 1_{m-T\le k\le l+T} c_{k-l} c_{k-m} & (2.23)\\ a^{3,n}_{l,m} &= 2\sum_{k=1}^{n} 1_{k<l-T} c_{k-l} c_{k-m} & (2.24)\\ a^{4,n}_{l,m} &= \sum_{k=1}^{n} 1_{k<m-T} 1_{l-T\le k\le l+T} c_{k-l} c_{k-m}, & (2.25)\end{aligned}$$

we find that

$$\begin{aligned} E\left[(S^{(i)}_n)^2\right] &= \sum_{l_1=-\infty}^{\infty}\sum_{m_1=l_1+1}^{\infty} a^{i,n}_{l_1,m_1} \sum_{l_2=-\infty}^{\infty}\sum_{m_2=l_2+1}^{\infty} a^{i,n}_{l_2,m_2} E\left[\xi_{l_1}\xi_{m_1}\xi_{l_2}\xi_{m_2}\right] \\ &= \sum_{l_1=-\infty}^{\infty}\sum_{m_1=l_1+1}^{\infty} a^{i,n}_{l_1,m_1} \sum_{l_2=-\infty}^{\infty}\sum_{m_2=l_2+1}^{\infty} a^{i,n}_{l_2,m_2} \delta_{l_1,l_2}\delta_{m_1,m_2} \\ &= \sum_{l=-\infty}^{\infty}\sum_{m=l+1}^{\infty} \left(a^{i,n}_{l,m}\right)^2 & (2.26)\end{aligned}$$

and for $n_r \le n < o < n_{r+1}$

$$E\left[(S^{(i)}_o - S^{(i)}_n)^2\right] = \sum_{l=-\infty}^{\infty}\sum_{m=l+1}^{\infty} \left(a^{i,o}_{l,m} - a^{i,n}_{l,m}\right)^2 \qquad (2.27)$$

for $i = 2, 3, 4$. Using a change of variables and the Beta distribution pdf, we have that

$$\begin{aligned} \sum_{l=k+1}^{j-1} c_{j-l} c_{k-l} &\overset{j,k}{\ll} \int_k^j (j-t)^{-\sigma} (t-k)^{-\sigma}\, dt \\ &= (j-k)^{1-2\sigma} \underbrace{\int_0^1 (1-s)^{-\sigma} s^{-\sigma} ds}_{B(1-\sigma,1-\sigma)} \overset{j,k}{\ll} (j-k)^{1-2\sigma}\,. & (2.28)\end{aligned}$$

Step 6: Apply $S^{(1)}$-procedure for convergence of large c terms $\frac{S_n^{(2)}}{n^{\frac{1}{p}}}$. Using (2.28) and integral approximation, one has for $n \in [n_r, n_{r+1})$

$$\begin{aligned}
& E\left[(S_n^{(2)})^2\right] - 4\sum_{k=1}^{n}\sum_{m>l} 1_{k-T\le m\le k+T}\cdot 1_{k-T\le l\le k+T}\, c_{k-l}^2 c_{k-m}^2 \\
=\ & 8\sum_{j>k}\sum_{m>l} 1_{j-T\le m\le k+T}\cdot 1_{j-T\le l\le k+T}\, c_{j-l}c_{j-m}c_{k-l}c_{k-m} \\
\le\ & 4\sum_{j>k}\left|\sum_{l=j-T}^{k+T} c_{j-l}c_{k-l}\right|^2 \\
\le\ & 4\sum_{k=1}^{n}\sum_{j=k+1}^{n\wedge(k+2T)}\left|2c_{j-k} + \sum_{l=j-T}^{k-1} c_{j-l}c_{k-l} + \sum_{l=k+1}^{j-1} c_{j-l}c_{k-l} + \sum_{l=j+1}^{k+T} c_{j-l}c_{k-l}\right|^2 \\
\overset{n}{\ll} & \sum_{k=1}^{n}\sum_{j=k+1}^{k+2T}\left[(j-k)^{-2\sigma} + (j-k)^{2-4\sigma} + (j-k)^{-2\sigma}T^{2-2\sigma}\right] \\
\overset{n}{\ll} &\ \ nl(n),
\end{aligned}$$

where $l\left(n\right) = \begin{cases} T^{3-4\sigma} = n_r^{\nu(3-4\sigma)} & \sigma < \frac{3}{4} \\ \log\left(T\right) = \nu\log(n_r) & \sigma = \frac{3}{4} \\ 1 & \sigma > \frac{3}{4} \end{cases}$. Hence,

$$E\left[(S_n^{(2)})^2\right] \overset{n}{\ll} nl\left(n\right) + \sum_{k=1}^{n}\left|\sum_{l=-T}^{T} c_l^2\right|^2 \overset{n}{\ll} nl\left(n\right). \tag{2.29}$$

Similarly, we have for $n_r \le n < o < n_{r+1}$ that

$$\begin{aligned}
E\left[\left(S_o^{(2)} - S_n^{(2)}\right)^2\right] \ \overset{o,n}{\ll}\ & \sum_{k=n+1}^{o}\left|\sum_{l=-T}^{T} c_l^2\right|^2 + \sum_{\substack{j,k=n+1 \\ j>k}}^{o}\left|\sum_{l=j-T}^{k+T} c_{j-l}c_{k-l}\right|^2 \\
\overset{o,n}{\ll}\ & \left(o-n\right)l\left(n\right). \qquad (2.30)
\end{aligned}$$

Therefore, it follows by Theorem 2.4.1 of Stout that

$$\begin{aligned} E\left[\max_{n_r \le n < o < n_{r+1}} (S_o^{(2)} - S_n^{(2)})^2\right] &\overset{r}{\ll} \left(\frac{\log(2n_r)}{\log 2}\right)^2 (n_{r+1} - n_r) l(n_{r+1}) \\ &\overset{r}{\ll} r^2 n_r l(n_r). \end{aligned} \tag{2.31}$$

Combining (2.29) with $n = n_r$ and (2.31), one has that

$$E\left[\sum_{r=0}^{\infty} \max_{n_r \le n < n_{r+1}} \left(\frac{S_n^{(2)}}{n^{\frac{1}{p}}}\right)^2\right] \ll \sum_{r=0}^{\infty} r^2 n_r^{1-\frac{2}{p}} l(n_r) < \infty, \tag{2.32}$$

provided $1 + \nu(3 - 4\sigma) \vee 0 < \frac{2}{p}$ (i.e. $p < \frac{2}{1+\nu(3-4\sigma)}$ when $\sigma < \frac{3}{4}$ and $p < 2$ when $\sigma \ge \frac{3}{4}$, both of which are true). It follows that $\lim\limits_{n\to\infty} \frac{S_n^{(2)}}{n^{\frac{1}{p}}} = 0$ a.s.

Step 7: Apply $S^{(1)}$-procedure for convergence of small c terms $\frac{S_n^{(3)}}{n^{\frac{1}{p}}}$.

$$\begin{aligned} & E\left[(S_n^{(3)})^2\right] \\ =& 8 \sum_{j>k} \sum_{m>l} 1_{j+T<l} \cdot 1_{k+T<l} c_{j-l} c_{j-m} c_{k-l} c_{k-m} \\ +& 4 \sum_{k=1}^{n} \sum_{m>l} 1_{k+T<l} c_{k-l}^2 c_{k-m}^2 \\ \le& 4 \sum_{j>k} \left| \sum_{l=j+T+1}^{\infty} c_{j-l} c_{k-l} \right|^2 + 2 \sum_{k=1}^{n} \left| \sum_{l=k+T+1}^{\infty} c_{k-l}^2 \right|^2 \\ \overset{n}{\ll}& \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \left| \int_{j+T}^{\infty} (t-j)^{-\sigma} (t-k)^{-\sigma} \, dt \right|^2 + \sum_{k=1}^{n} \left| \int_{k+T}^{\infty} (t-k)^{-2\sigma} \, dt \right|^2 \\ \overset{n}{\ll}& \sum_{k=1}^{n} \left(\sum_{j=k+1}^{n} \left| \int_{T}^{\infty} t^{-2\sigma} dt \right|^2 + \left| \int_{T}^{\infty} t^{-2\sigma} dt \right|^2 \right) \\ \overset{n}{\ll}& n^2 T^{2-4\sigma}. \end{aligned} \tag{2.33}$$

Similarly, we have for $n_r \leq n < o < n_{r+1}$ that

$$E\left[\left(S_o^{(3)} - S_n^{(3)}\right)^2\right] \overset{o,n,r}{\ll} (o-n)\, oT^{2-4\sigma} \overset{o,n,r}{\ll} (o-n)\, n_{r+1}^{1+\nu(2-4\sigma)}. \qquad (2.34)$$

Therefore, it follows by Theorem 2.4.1 of Stout that

$$\begin{aligned} E\left[\max_{n_r \leq n < o < n_{r+1}} \left(S_o^{(3)} - S_n^{(3)}\right)^2\right] &\overset{r}{\ll} \left(\frac{\log(2n_r)}{\log 2}\right)^2 (n_{r+1} - n_r) n_{r+1}^{1+\nu(2-4\sigma)} \\ &\overset{r}{\ll} r^2 n_r^{2+\nu(2-4\sigma)}. \qquad (2.35) \end{aligned}$$

Combining (2.33) with $n = n_r$ and (2.35), one has

$$E\left[\sum_{r=0}^{\infty} \max_{n_r \leq n < n_{r+1}} \left(\frac{S_n^{(3)}}{n^{\frac{1}{p}}}\right)^2\right] \ll \sum_{r=0}^{\infty} r^2 n_r^{2+\nu(2-4\sigma)-\frac{2}{p}} < \infty, \qquad (2.36)$$

provided $p < \frac{1}{1+\nu(1-2\sigma)}$, which is the given condition, so $\lim_{n\to\infty} \frac{S_n^{(3)}}{n^{\frac{1}{p}}} = 0$ a.s..

It is notable that condition on p, $p < \frac{2}{1+\nu(3-4\sigma)}$, in step 6 gets more stringent when $\nu > 1$ and the same is true for condition on p, $p < \frac{1}{1+\nu(1-2\sigma)}$, in step 7 when $\nu < 1$, so the best choice that raises the same condition on p is when $\nu = 1$. Hence, we will have to satisfy $p < \frac{1}{1-2\sigma}$ in either cases.

Step 8: Apply $S^{(1)}$-procedure for convergence of mixed terms $\frac{S_n^{(4)}}{n^{\frac{1}{p}}}$.

Finally, we note

$$\begin{aligned} & E\left[(S_n^{(4)})^2\right] \\ = & \sum_{k=1}^{n} \sum_{m=k+T+1}^{\infty} c_{k-m}^2 \sum_{l=k-T}^{l=k+T} c_{k-l}^2 + 2\sum_{k=1}^{n} \sum_{j=k+1}^{k+2T} \sum_{m=j+T+1}^{\infty} c_{j-m} c_{k-m} \sum_{l=j-T}^{k+T} c_{j-l} c_{k-l} \\ \overset{n}{\ll} & \sum_{k=1}^{n} \left\{ T^{1-2\sigma} + \sum_{j=k+1}^{k+2T} T^{1-2\sigma} \left[(j-k)^{-\sigma} + (j-k)^{1-2\sigma} + (j-k)^{-\sigma} T^{1-\sigma}\right] \right\} \\ \overset{n}{\ll} & nT^{3-4\sigma}. \end{aligned}$$

Similarly, we have for $n_r \leq n < o < n_{r+1}$ that

$$E\left[\left(S_o^{(4)} - S_n^{(4)}\right)^2\right] \overset{o,n}{\ll} (o-n)\, T^{3-4\sigma}.$$

Therefore, it follows by $\nu = 1$ and Theorem 2.4.1 of Stout that

$$E\left[\max_{n_r \leq n < o < n_{r+1}} \left(S_o^{(4)} - S_n^{(4)}\right)^2\right] \overset{r}{\ll} \left(\frac{\log(2n_r)}{\log 2}\right)^2 (n_{r+1} - n_r) n_{r+1}^{3-4\sigma} \overset{r}{\ll} r^2 n_r^{4-4\sigma}.$$

Combining these two equations, one has

$$E\left[\sum_{r=0}^{\infty} \max_{n_r \leq n < n_{r+1}} \left(\frac{S_n^{(4)}}{n^{\frac{1}{p}}}\right)^2\right] \;\; \ll \;\; \sum_{r=0}^{\infty} r^2 n_r^{(4-4\sigma)-\frac{2}{p}} < \infty, \qquad (2.37)$$

provided $p < \frac{1}{2-2\sigma}$, which is true. It follows that $\lim\limits_{n\to\infty} \frac{S_n^{(4)}}{n^{\frac{1}{p}}} = 0$ a.s. □

2.4.2 Proof of Theorem 2.3

Without loss of generality we assume $1 < \alpha < 2$.

Step 1: Reduce to continuous $\{(\xi_l, \overline{\xi}_l)\}$.

Let $\{(U_l)\}_{l\in\mathbb{Z}}$ be independent $[-1,1]$-uniform random variables that are independent of everything and set $\overline{U}_l = U_l$ for all l. Then, we have that

$$\begin{aligned} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (d_k - d) \;&=\; \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} \sum_{l,m=-\infty}^{\infty} c_{k-l}\overline{c}_{k-m} \left((\xi_l + U_l)(\overline{\xi}_m + \overline{U}_m) - d - \frac{2}{3}\right) \\ &- \;\frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} \sum_{l,m=-\infty}^{\infty} c_{k-l}\overline{c}_{k-m} \left(\xi_l \overline{U}_m + U_l \overline{\xi}_m + U_l \overline{U}_m - \frac{2}{3}\right). (2.38) \end{aligned}$$

However,

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} \sum_{l,m=-\infty}^{\infty} c_{k-l}\overline{c}_{k-m} \left(\xi_l \overline{U}_m + U_l \overline{\xi}_m + U_l \overline{U}_m - \frac{2}{3}\right) = 0 \qquad (2.39)$$

by Theorem 2.8. Moreover, $\xi_1 + U_1, \overline{\xi}_1 + \overline{U}_1$ have the same moment and tail probability bounds as $\xi_1, \overline{\xi}_1$. Hence, without loss of generality, we can assume $\xi_l, \overline{\xi}_m$ are continuous random variables, which will be important for the truncation to follow in Step 4.

Step 2: Handle off-diagonal sum as previous proof since unaffected by heavy tails.

Suppose $S_n^{(2)}$, $S_n^{(3)}$ and $S_n^{(4)}$ are defined as in (2.15-2.17). Then, we know that

$$\lim_{n\to\infty} \frac{S_n^{(2)}}{n^{\frac{1}{p}}} = \lim_{n\to\infty} \frac{S_n^{(3)}}{n^{\frac{1}{p}}} = \lim_{n\to\infty} \frac{S_n^{(4)}}{n^{\frac{1}{p}}} = 0 \text{ a.s.},$$

provided $p < \frac{1}{2-\sigma-\overline{\sigma}}$ by the proof of Theorem 2.8.

Step 3: Reduce $\xi_l\overline{\xi}_l$ (in diagonal sum) to non-negative with single atom at 0.

Noting

$$\begin{aligned}&\sum_{l=-\infty}^{\infty} c_{k-l}\overline{c}_{k-l}(\xi_l\overline{\xi}_l - E[\xi_l\overline{\xi}_l])\\ =&\sum_{l=-\infty}^{\infty} c_{k-l}\overline{c}_{k-l}((\xi_l\overline{\xi}_l)^+ - E[(\xi_l\overline{\xi}_l)^+]) - \sum_{l=-\infty}^{\infty} c_{k-l}\overline{c}_{k-l}((\xi_l\overline{\xi}_l)^- - E[(\xi_l\overline{\xi}_l)^-]),\end{aligned} \quad (2.40)$$

we only have to consider the case where $\xi_l\overline{\xi}_l \geq 0$ for the remainder of the proof. Moreover, insomuch as the proof of the general case only differs cosmetically from the notationally-simpler case where $\overline{\xi}_l = \xi_l$, $E[\xi_1^2] = 1$ and $\overline{c}_l = c_l = \begin{cases} 1 & l = 0 \\ |l|^{-\sigma} & l \neq 0 \end{cases}$, we only provide the proof of the later for which the long-range dependence constraint becomes $p < \frac{1}{2-2\sigma}$. We will however indicate the most significant changes that would be needed for the general case.

Step 4: Divide diagonal terms into zero-mean truncated (i.e. bounded) and remainder pieces.

Let $\kappa > 0$. Fix $u_r^+ = n_r^{\frac{\kappa}{2-\alpha}}$ to find

$$2\int_0^{u_r^+} P(\xi_1^2 > s)sds \overset{r}{\ll} 2\int_0^{u_r^+} ss^{-\alpha}ds \overset{r}{\ll} n_r^\kappa \;\; \forall\, r = 1, 2, ... \tag{2.41}$$

Now, by defining

$$\begin{cases} \overline{\zeta}_i = \overline{\zeta}_i^r = (\xi_i^2 \wedge u_r^+) - \vartheta_i, \text{ where } \vartheta_i \doteq \int_0^{u_r^+} P(\xi_i^2 > s)ds \le 1, \\ \tilde{\zeta}_i = \tilde{\zeta}_i^r = \xi_i^2 - 1 - \overline{\zeta}_i^r, \end{cases} \tag{2.42}$$

we find that

$$E[\overline{\zeta}_i] = \int_0^{u_r^+} P(\xi_i^2 > t)dt - \int_0^{u_r^+} P(\xi_i^2 > t)dt = 0, \tag{2.43}$$

so both $\overline{\zeta}_i$ and $\tilde{\zeta}_i$ are zero mean, and by (2.41)

$$\begin{aligned} E[|\overline{\zeta}_1|^2] &= E|\xi_1^2 \wedge u_r^+|^2 - \left(\int_0^{u_r^+} P(\xi_1^2 > t)dt\right)^2 \\ &= 2\int_0^{u_r^+} P(\xi_1^2 > s)sds - \left(\int_0^{u_r^+} P(\xi_1^2 > t)dt\right)^2 \\ &\overset{r}{\ll} n_r^\kappa \;\; \forall\, r = 1, 2, ... \end{aligned} \tag{2.44}$$

(In the general case, we note that $\xi_1\overline{\xi}_1$ is non-negative and of continuous distribution on $(0, \infty)$ so $E[\xi_1\overline{\xi}_1 \wedge u_r^+] = \int_0^{u_r^+} P(\xi_1\overline{\xi}_1 > s)ds$ as required. We also have $\tilde{\zeta}_i^r = \xi_i\overline{\xi}_i - E[\xi_i\overline{\xi}_i] - \overline{\zeta}_i^r$.)

Step 5: Moment Bound for truncated using the proof of Theorem 2.8.

Noting $\{\overline{\zeta}_i\}$ are i.i.d. with $E[\overline{\zeta}_1] = 0$ and $E[\overline{\zeta}_1^2] < \infty$ and defining

$$S_n^{(1)} = \sum_{k=1}^{n} \sum_{l=-\infty}^{\infty} c_{k-l}^2 \overline{\zeta}_l, \tag{2.45}$$

one finds from (2.22) in the proof of Theorem 2.8 that

$$E\left[\max_{n_r \le n < n_{r+1}} \left(S_n^{(1)}\right)^2\right] \le E|\overline{\zeta}_1|^2 r^2 n_r. \tag{2.46}$$

Hence, it follows by (2.44) that

$$E\left[\max_{n_r \le n < n_{r+1}} \left(S_n^{(1)}\right)^2\right] \overset{r}{\ll} r^2 n_r^{1+\kappa}. \tag{2.47}$$

Step 6: Moment Bound for remainder using Doob's inequality.

Turning to the $\tilde{\zeta}_i^r$ and using the formula

$$E[g(X)] = \int_0^\infty g'(t)P(X > t)dt - \int_{-\infty}^0 g'(t)P(X < t)dt, \tag{2.48}$$

one has by our tail probability bounds that the non-negative part of $\tilde{\zeta}_1$ satisfies

$$\begin{aligned} E|\tilde{\zeta}_1^+|^\tau &= \tau \int_0^\infty s^{\tau-1} P(\xi_1^2 > u_r^+ + s + 1 - \vartheta_1) ds \\ &\le \tau \int_0^\infty s^{\tau-1} P(\xi_1^2 > u_r^+ + s) ds \text{ since } \vartheta_1 \le 1 \\ &\overset{r}{\ll} \int_{u_r^+}^\infty (s - u_r^+)^{\tau-1} s^{-\alpha} ds \\ &\le \int_{u_r^+}^{2u_r^+} (s - u_r^+)^{\tau-1} ds (u_r^+)^{-\alpha} + \int_{2u_r^+}^\infty (s - u_r^+)^{\tau-\alpha-1} ds \\ &\overset{r}{\ll} (u_r^+)^{\tau-\alpha} \overset{r}{\ll} n_r^{\frac{\kappa(\tau-\alpha)}{2-\alpha}}, \end{aligned} \tag{2.49}$$

for $1 < \tau < \alpha$. Therefore, it follows by Jensen's inequality and Doob's L_p

inequality that

$$\begin{aligned}
E^{\frac{1}{\tau}}\left[\sup_{n_r\le n<n_{r+1}}\left|\sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}c_l^2\tilde{\zeta}_{k-l}^r\right|^{\tau}\right] &\le E^{\frac{1}{\tau}}\left[\left|\sum_{l=-\infty}^{\infty}c_l^2\sup_{n_r\le n<n_{r+1}}\left|\sum_{k=1}^{n}\tilde{\zeta}_{k-l}^r\right|\right|^{\tau}\right]\\
&\overset{r}{\ll}\sum_{l=-\infty}^{\infty}c_l^2E^{\frac{1}{\tau}}\left[\sup_{n_r\le n<n_{r+1}}\left|\sum_{k=1}^{n}\tilde{\zeta}_{k-l}^r\right|^{\tau}\right]\\
&\overset{r}{\ll}\sum_{l=-\infty}^{\infty}c_l^2E^{\frac{1}{\tau}}\left[\left|\sum_{k=1}^{n_{r+1}-1}\tilde{\zeta}_{k-l}^r\right|^{\tau}\right]\\
&\overset{r}{\ll} n_r\|\tilde{\zeta}_1^r\|_{\tau}, \qquad (2.50)
\end{aligned}$$

so by (2.49, 2.50)

$$E\left[\sup_{n_r\le n<n_{r+1}}\left|\sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}c_l^2\tilde{\zeta}_{k-l}^r\right|^{\tau}\right]\overset{r}{\ll} n_r^{\tau-\frac{\kappa(\alpha-\tau)}{2-\alpha}}. \qquad (2.51)$$

Step 7: Use Truncation and Error Term bounds with Borel-Cantelli for convergence.

Combining (2.47) and (2.51), one has that

$$\begin{aligned}
&P\left(\sup_{n_r\le n<n_{r+1}}\left|\sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}c_l^2\zeta_{k-l}\right|>2\epsilon n_r^{\frac{1}{p}}\right)\\
\le\ & \frac{E\left[\sup_{n_r\le n<n_{r+1}}\left|\sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}c_l^2\overline{\zeta}_{k-l}^r\right|^2\right]}{\epsilon^2 n_r^{\frac{2}{p}}}+\frac{E\left[\sup_{n_r\le n<n_{r+1}}\left|\sum_{k=1}^{n}\sum_{l=-\infty}^{\infty}c_l^2\tilde{\zeta}_{k-l}^r\right|^{\tau}\right]}{\epsilon^{\tau}n_r^{\frac{\tau}{p}}}\\
\overset{r}{\ll}\ & r^2n_r^{1+\kappa-\frac{2}{p}}+n_r^{\tau-\frac{\kappa(\alpha-\tau)}{2-\alpha}-\frac{\tau}{p}}\\
\overset{r}{\ll}\ & r^2n_r^{1-\frac{\alpha}{p}}+n_r^{\tau-\frac{\alpha}{p}}, \qquad (2.52)
\end{aligned}$$

by letting $\kappa = \frac{2-\alpha}{p}$. Hence, if $\tau \in \left(1, \frac{\alpha}{p}\right)$, then

$$\sum_{r=1}^{\infty} P\left(\sup_{n_r \leq n < n_{r+1}} \left|\sum_{k=1}^{n} \sum_{l=-\infty}^{\infty} c_l^2 \zeta_{k-l}\right| > 2\epsilon n_r^{\frac{1}{p}}\right) < \infty, \tag{2.53}$$

under our heavy-tail condition $p < \alpha$ and

$$n^{-\frac{1}{p}} \sum_{k=1}^{n} \sum_{l=-\infty}^{\infty} c_l^2 \zeta_{k-l} \to 0 \quad \text{a.s.}, \tag{2.54}$$

by Borel-Cantelli. The proof is complete. □

Bibliography

[1] Avram, F. and Taqqu, M.S. (1987). *Generalized powers of strongly dependent random variables.* Ann. Probab., **vol.** 15, pp. 767-775.

[2] Davis, R.A. and Resnick, S.I. (1986). *Limit theory for the sample covariance and correlation functions of moving averages.* Ann. Statist., **vol.** 14, pp. 533-558.

[3] Dobrushin, R.L. and Major, P. (1979). *Non-central limit theorems for non-linear functions of Gaussian fields.* Z. Wahrscheinlichkeitstheorie Verw. Geb., **vol.** 50, pp. 27-52.

[4] Giraitis L. and Surgailis, D. (1986). *Multivariate Appell polynomials and the central limit theorem.* In E. Eberlein and M. S. Taqqu, (eds.), Dependence in Probability and Statistics, Birkhäuser, Boston, pp. 21-71.

[5] Giraitis L. and Surgailis, D. (1989). *Limit theorem for polynomials of linear process with long-range dependence.* Lith. Math. J., **vol.** 29, pp. 128-145.

[6] Horváth, L. and Kokoszka, P. (2008). *Sample autocovariances of long-memory time series.* Bernoulli, **vol.** 14, pp. 405-418.

[7] Kouritzin, M.A. and Sadeghi, S. (2015). *Convergence Rates and Decoupling in Linear Stochastic Approximation Algorithms.* SIAM Journal on Control and Optimization, **vol.** 53-3, pp. 1484-1508.

[8] Kouritzin, M.A. (1996). *On the convergence of linear stochastic approximation procedures.* IEEE Trans. Inform. Theory, **vol.** 42, pp. 1305-1309.

[9] Kouritzin, M.A. (1996). *On the interrelation of almost sure invariance principles for certain stochastic adaptive algorithms and for partial sums of random variables.* J. Theoret. Probab., **vol.** 9, No. 4, pp. 811-840.

[10] Kouritzin, M.A. (1995). *Strong approximation for cross-covariances of linear variables with long-range dependence.* Stochastic Processes Appl. **vol.** 60, pp. 343-353.

[11] Louhchi, S. and Soulier, P. (2000). *Marcinkiewicz-Zegmond Strong Laws for Infinite Variance Time Series.* Statistical Inference for Stochastic Processes, **vol.** 3, pp. 31-40.

[12] Mandelbrot, B. and Wallis J. (1968). *Noah, Joseph and operational hydrology.* Water Resources Research, **vol.** 4, pp. 909-918.

[13] Mandelbrot, B. (1972). *Statistical methodology for non-periodic cycles: from the covariance to R/S analysis.* Ann. Econ. and Social Measurement, **vol.** 1, pp. 259-290.

[14] Rosenblatt, M. (1961). *Independence and dependence.* Proc. 4th Berkeley Symp. Math. Statist. Probab., pp. 411-443.

[15] Stout, W.F. (1974). *Almost Sure Convergence.* Academic Press Inc., pp. 126.

[16] Surgailis, D. (1982). *Zones of attraction of self-similar multiple integrals.* Lithuanian Math. J., **vol.** 22, pp. 327-340.

[17] Surgailis, D. (2004). *Stable limits of sums of bounded functions of long-memory moving averages with finite variance.* Bernoulli, **vol.** 10, pp. 327-355.

[18] Taqqu, M.S. (1979). *Convergence of integrated processes of arbitrary Hermite rank.* Z. Wahrscheinlichkeitstheorie Verw. Geb., **vol.** 50, pp. 53-83.

[19] Vaiciulis, M. (2003). *Convergence of sums of Appell polynomials with infinite variance.* Lithuanian Math. J., **vol.** 43, pp. 80-98.

[20] Varotsos, C. and Kirk-Davidoff, D. (2006). *Long-memory processes in global ozone and temperature variations at the region* 60^0 S-60^0 N*.* Atmospheric Chemistry and Physics, **vol.** 6, pp. 4093-4100.

[21] Wu, W.B. and Min, W. (2005). *On linear processes with dependent innovations.* Stochastic Processes and their Applications, **vol.** 115, pp. 939-958.

[22] Wu, W.B., Huang, Y. and Zheng, W. (2010). *Covariances Estimation for Long-Memory Processes.* Adv. Appl. Prob., **vol.** 42, pp. 137-157.

Chapter 3

Convergence Rates and Decoupling in Linear Stochastic Approximation Algorithms[*]

3.1 Introduction

Linear stochastic approximation algorithms have found widespread application in parameter estimation, adaptive machine learning, signal processing, econometrics and pattern recognition (see, e.g., [1], [3], [9], [26] and [32]). Consequently, their asymptotic rates of almost sure and r^{th}-mean convergence as well as invariance and large deviation principles are of utmost importance (see e.g., [6], [11], [17], [18], [21], [22], [24], [34] and [36]). For motivation, suppose $\{x_k, k = 1, 2, \cdots\}$ and $\{y_k, k = 2, 3, \cdots\}$ are second order $\mathbb{R}^d-$ and

[*] A version of this chapter has been published. Kouritzin, M.A. and Sadeghi, S. (2015). *Convergence Rates and Decoupling in Linear Stochastic Approximation Algorithms.* SIAM Journal on Control and Optimization, **vol.** 53-3, pp. 1484-1508.

$\mathbb{R}-$valued stochastic processes, defined on some probability space $(\Omega, \mathcal{F}, P)$, that satisfy

$$y_{k+1} = x_k^T h + \epsilon_k, \qquad \forall k = 1, 2, \ldots, \tag{3.1}$$

where h is an unknown d-dimensional parameter or weight vector of interest and ϵ_k is a noise sequence. One often wants to find the value of h that minimizes the mean-square error $h \to E|y_{k+1} - x_k^T h|^2$. This *best* h is given by $h = A^{-1}b$, where $A = E(x_k x_k^T)$ and $b = E(y_{k+1} x_k)$, assuming the expectations exist, wide-sense stationarity conditions and that A is positive definite. However, we often do not know the joint distribution of (x_k, y_{k+1}) nor have the necessary stationarity but instead estimate h using a linear algorithm of the form:

$$h_{k+1} = h_k + \mu_k(b_k - A_k h_k), \tag{3.2}$$

where μ_k is the k^{th} step size (often of the form $\mu_k = k^{-\chi}$ for some $\chi \in \left(\frac{1}{2}, 1\right]$) and

$$A_k = \frac{1}{N} \sum_{l=\max\{k-N+1,1\}}^{k} x_l x_l^T, \text{ and } b_k = \frac{1}{N} \sum_{l=\max\{k-N+1,1\}}^{k} y_{l+1} x_l \tag{3.3}$$

for some $N \in \mathbb{N}$, are random sequences of symmetric, positive-semi-definite matrices and vectors respectively. Most often $N = 1$ so $A_k = x_k x_k^T$ and $b_k = y_{k+1} x_k$. More information on stochastic approximation can be found in e.g. [5], [8], [10], [13], [17], [25], [30] and [37], which provide examples and motivation for our work. However, our work is easily differentiated from these. Delyon [8], for example, focuses on non-linear stochastic approximation algorithms, treating linear examples the same as non-linear ones. (In Section 4.2.2 he uses linear algorithm approximation but with a constant deterministic matrix $A_k = A$ in our notation.) Delyon's work handles important applications. However, his A-stable and (A, B) Conditions are usually harder to verify than our Marcinkiewicz Strong Law of Large Numbers (MSLLN) conditions (given below) in the (unbounded, random A_k) linear case, he does not supply almost sure rates of convergence, his theorems are geared to martingale-increment-plus-decreasing-perturbation noise and he often assumes fourth order moments. We are motivated by (but not restricted to) the common setting where $X_k^T = (x_k^T, y_{k+1})$ is a (multivariate) linear process

$$X_k = \sum_{l=-\infty}^{\infty} C_{k-l}\Xi_l. \tag{3.4}$$

Matrix sequence (C_l) can decay slowly enough (as $|l| \to \infty$) for long-range dependence (LRD) while $\{\Xi_l\}$ can have heavy tails (HT), so $E|b_k|^2 = \infty$ and/or $E|A_k|^2 = \infty$. Even in the lighter tail, short-range dependence case our two-sided linear process example $\{x_k\}$ is not a martingale. Moreover; long-range dependence and heavy tails; exhibited in many network [19], financial and paleoclimatic data sets for example; voids the usual mixing and moment conditions. We focus on one-step versus Polyak-Ruppert's two-step averaging algorithms but handle heavy tails and long-range dependence, deriving a surprising decoupling. This means that the optimal convergence rate of (3.2) is affected by either the heavy tails or the long-range dependence, whichever is worse, but not both. This contrasts the rate for partial sums of long-range dependent, heavily-tailed random variables, which is degraded twice (see e.g. Theorem 3.4).

Step size μ_k has a direct effect on the convergence rate and algorithm effectiveness (see, e.g [12], [15] and references cited therein). Consider the extreme cases. In the homogeneous, deterministic setting, i.e. $A_k = A$ and $b_k = b$, (3.2) can solve the linear equation $Ah = b$ when matrix inversion of A is ill-conditioned. In this case, a constant gain $\mu_k = \epsilon$ is best: Since $b = Ah$, we have $h_{k+1} = h_k - \epsilon A(h_k - h)$, so $h_n - h = (I - \epsilon A)^{n-1}(h_1 - h)$ and $h_n \to h$ geometrically, provided ϵ is small enough that the eigenvalues of $I - \epsilon A$ are within the unit disc. Conversely, in the presence of persistent noise, decreasing step sizes are required for the convergence $h_n \to h$. Existing results show that the best possible almost-sure rate of convergence is $|h_n - h| = O\left(\sqrt{n^{-1}\log\log(n)}\right)$, implied by the law of the iterated logarithm, and that this rate is only attainable when $\mu_k = \frac{1}{k}$, second moments of A_k, b_k exist and there is no long-range dependence. (These claims follow from the almost-sure invariance principle in Kouritzin [22].)

Herein, we handle all gains, long-range dependence and heavy tails, addressing the optimal rate of convergence by establishing results akin to the MSLLN, namely $n^\gamma |h_n - h| \to 0$ a.s. (i.e. $|h_n - h| = o(n^{-\gamma})$), for all $\gamma < \gamma_0(\chi) \doteq \chi - M$. M is called the *Marcinkiewicz threshold* in the sequel and is defined by

$$M \doteq \inf\left\{\frac{1}{m} : \lim_{n\to\infty} \frac{1}{n^{\frac{1}{m}}} \sum_{k=1}^{n}(A_k - A) = 0, \lim_{n\to\infty} \frac{1}{n^{\frac{1}{m}}} \sum_{k=1}^{n}(b_k - b) = 0 \text{ a.s.}\right\}. \quad (3.5)$$

Usually, we expect $M \in (\frac{1}{2}, 1]$, due to Strong Law of Large Numbers and Central Limit Theorem in the light-tail, short-range-dependence case but when there is LRD and/or HT M generally cannot approach $\frac{1}{2}$. When $\{(x_k^T, y_{k+1})^T : k \in \mathbb{Z}\}$ is a linear process as in (3.4), it is shown in [20] that $M = \frac{1}{\alpha} \vee (2 - 2\sigma)$

with $\alpha \doteq \sup\{a \leq 2 : \sup_{t\geq 0} t^a P(|\Xi_1|^2 > t) < \infty\}$ and $\sigma \doteq \sup\{s \in (\frac{1}{2}, 1] : \sup_l |l|^s \|C_l\| < \infty\}$. Hence, $\gamma < \gamma_0(\chi) \doteq (\chi - \frac{1}{\alpha}) \wedge (\chi + 2\sigma - 2)$. Here, $\alpha \in (1, 2]$ is a heavy-tail parameter with $\alpha = 2$ indicating non-heavy tails and $\sigma \in \left(\frac{1}{2}, 1\right]$ is a long-range dependence parameter with $\sigma = 1$ indicating the minimal amount of long-range dependence.

In classical applications the best theoretical convergence rate is attained when $\chi = 1$ corresponding to $\gamma_0(\chi) = \frac{1}{2}$. However, this rate knowledge can lead to erroneous conclusions as the algorithm often performs better with $\mu_k = k^{-\chi}$ for some $\chi < 1$ or even constant gain (see [23]) than with $\mu_k = \frac{1}{k}$. How might one explain this apparent paradox? First of all, these simple rate-of-convergence results do not account for the possibility of exploding constants, i.e. if $h_k^{(\chi)}$ denotes the solution of the algorithm (3.2) with $\mu_k = k^{-\chi}$, then $|h_n(\chi) - h| = D^\chi n^{-\gamma(\chi)}$ for all $\gamma(\chi) < \gamma_0(\chi)$. However, this D^χ often increases rapidly as $\chi \nearrow 1$ so the *observed* convergence may be fastest for some $\chi < 1$. Secondly, a higher value of χ is worse for forgetting a poor initial guess h_0 of h since you move further and further from the geometric convergence mentioned above as $\chi \to 1$.

Our approach is to transfer the MSLLN from the partial sums of a linear algorithm's coefficients to its solution. In other words, we establish the almost sure rates of convergence $n^\gamma |h_n - h| \to 0$ a.s., for the algorithm

$$h_{k+1} = h_k + \frac{1}{k^\chi}(b_k - A_k h_k) \;\; \forall\; k = 1, 2, 3, ... \tag{3.6}$$

with $\chi \in (0, 1)$, assuming only

$$\lim_{n\to\infty} \frac{1}{n^\chi} \sum_{k=1}^{n} (A_k - A) = 0 \;\text{ and }\; \lim_{n\to\infty} \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (b_k - A_k h) = 0 \;\text{ a.s.} \tag{3.7}$$

for some $\gamma \in [0, \chi)$, which can be implied by e.g.

$$\lim_{n\to\infty} \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (A_k - A) = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (b_k - b) = 0 \quad \text{a.s.}, \tag{3.8}$$

where $Ah = b$. When $\chi - \gamma \in (\frac{1}{2}, 1]$, these conditions can be verified by the MSLLN under a variety of conditions, which we study using the specific structure of A_k and b_k in Section 3.4.

In addition to rates of convergence, our results show that convergence ($h_k \to h$) in (3.6) takes place provided that $\chi \in (M, 1)$. All this suggests that more quickly decreasing gains like $\mu_k = \frac{1}{k^\chi}$ with χ near 1 should be used in very heavy-tailed or long-range dependent settings. Conversely, slowly deceasing gains like $\mu = \frac{1}{k^\chi}$ with smaller χ might work well in lighter-tailed, short-range-dependent situations. Our simulations in Section 3.5 show that the smallest normalized error, $\frac{|h_n - h|}{|h_1 - h|}$, usually occurs for $\chi \in (M, 1]$ and the most commonly used choice $\chi = 1$ is most appropriate in very heavy-tailed or long-range-dependent settings (where M is close to 1) or very long runs. In other words, a slower decreasing gain usually gets you close to the true parameters h more quickly unless the coefficients have a high probability of differing significantly from their means.

Let us consider what is new in terms of our theoretical results. The idea of inferring convergence and rates of convergence results for linear algorithms (3.2) from like convergence and rates of convergence of its coefficients is not new. Indeed, it dates back at least to work done by one of the authors in 1994 and 1996 (see [21], [22] and [24]). The result [24] considered relatively general gain μ_k and achieved optimal rates of r^{th}-mean convergence. It has been proved in [22] that the solution of the linear algorithm (3.2) satisfies an almost sure invariance principle with respect to a limiting Gaussian process when $\mu_k = \frac{1}{k}$ and each A_k is symmetric under the minimal condition that the coefficients satisfy such an a.s. invariance principle. One could then immediately transfer functional laws of the iterated logarithm from the limiting Gaussian process back to the solution of the linear algorithm. Again assuming the "usual" conditions of A_k symmetry and $\mu_k = \frac{1}{k}$, Kouritzin [21] showed that the solution of the linear algorithm converges almost surely given that the coefficients do. While this result does not state rates of convergence, our current work in going from Proposition 3.1 to Theorem 3.1 within shows that almost-sure rate of convergence sometimes follow from convergence results for linear algorithms as a simple corollary.

There were many results (see, e.g. [13], [14] and [16]) that preceded those mentioned above and gave convergence or rates of convergence for linear algorithms. However, these results assumed a specific dependency structure and, thereby, were not generally applicable. More recently, some authors, e.g. [6], [8] and [34], have followed the path of transferring convergence and rates of convergence from partial sums of (the coefficient) random variables to the solutions of linear equations. Specifically, Tadić [34] transferred almost-sure rates of convergence, including those of the law-of-the-iterated-logarithm rate, from the coefficients to the linear algorithm in the non-symmetric-A_k, general-gain case. He does not develop a law of the iterated logarithm where one characterizes the limit points nor does he consider functional versions. Moreover, he imposes one of two sets of conditions (A and B in his notation). Conditions B ensure the gain $\mu_k \approx \frac{1}{k}$, so these results should be compared to prior results in [4] and [22], which imply stronger Strassen-type functional laws of the iterated logarithm. Tadić does not give any examples verifying his Conditions A where lesser rates are obtained.

It seems that we are the first to consider processes that are simultaneously heavy-tailed and long-range dependent in stochastic approximation.

The rest of this chapter is organized as follows. A motivational example is given next. The main theorems are formulated in Section 3.3. Then, Section 3.4 includes some background about the Marcinkiewicz Strong Law of Large Numbers for Partial Sums and a new MSLLN result for outer products of multivariate linear processes with LRD and HT. Experimental results are given in Section 3.5 and proof of main result (Theorem 3.1) is delayed until Section 3.6.

3.2 Example: Asymptotic Linear Observers by Adaptive Filtering

We refer to books Kushner and Yin [25], Ljung [27] and Soderstrom and Stoica [31] for standard vital applications in system identification, equalization, estimation, and adaptive control in stochastic systems. Rather than repeating these developments here, we just adapt a less-discussed, yet interesting application from Thanh, Yin and Wang [35]. They analyzed the convergence of double-indexed or triangular-array processes with mixing driving noises and random weights and established Marcinkiewicz Strong laws of large numbers and convergence rates for such problems (see Theorem 3.6 herein).

For motivation,Thanh, Yin and Wang considered the least square estimate of internal state of a multi-input-single-output linear-time-invariant system and showed that the estimation error is a special case of triangular-array processes with random weights and mixing driving noise. Alternatively to least squares, the internal states of linear observers can be estimated through adaptive filtering algorithms.

Consider the following linear time-invariant system operating near steady state

$$\begin{cases} \dot{X}(t) = AX(t) + Bu(t), \\ Y(t) = CX(t), \end{cases} \tag{3.9}$$

where $A \in \mathbb{R}^{m_0 \times m_0}$, $B \in \mathbb{R}^{m_0 \times m_1}$ and $C \in \mathbb{R}^{1 \times m_0}$ are known system matrices. We are interested about estimating the state $X(t)$. However, since we are operating near steady state $X(t) \approx h$ for some unknown h. $X(t)$, i.e. h, must be estimated through output Y. $Y(t)$ is measured only at a sequence of irregular (i.e. random) sampling time instants $\{t_k\}$ with measured values y_{k+1} corrupted by correlated noise $\{d_k\}$:

$$y_{k+1} = Y(t_k) + d_k. \tag{3.10}$$

(Irregular sampling time sequences $Y(t_k)$ are sometimes generated actively by input control or threshold adaptation under binary-valued sensors, or passively due to event-triggered sampling or low-resolution signal quantization.) The goal is to estimate the state $X(t)$ from information on the control input $u(t)$, $\{t_k\}$, and $\{y_{k+1}\}$, all of which are known or learnt in real time. The internal state in (3.9) satisfies

$$X(t_{k+1}) = e^{A(t_{k+1}-t_k)}X(t_k) + \int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bu(\tau)d\tau, \tag{3.11}$$

which we can not observe. Rather, we get access to the observations at the sampling time sequence $\{t_k, k = 1, ..., n\}$

$$\begin{aligned} y_{k+1} &= Ce^{A(t_{k+1}-t_k)}X(t_k) + C\int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bu(\tau)d\tau + d_k \\ &= x_k h + \nu_k + d_k, \end{aligned} \tag{3.12}$$

where $x_k = Ce^{A(t_{k+1}-t_k)}$ and $\nu_k = C\int_{t_k}^{t_{k+1}} e^{A(t_{k+1}-\tau)}Bu(\tau)d\tau$ can be built in real time from known system matrices and the observed sampling times. Letting $\epsilon_k = d_k + \nu_k$, we find (3.12) is the same as (3.1). Hence, the stochastic approximation algorithm can be used to recursively find the value of h that minimizes the mean-square error $h \to E|y_{k+1} - x_k^T h|^2$. h gives us the estimate of internal state of multi-input-single-output linear-time-invariant system X. In addition, if the estimated steady state h agrees with the calculated steady state value of X from the model itself, this supports the model and choice of system matrices.

If X were not close to steady state initially but the variation of $X(t_k)$ is rarely large between consecutive samples, then one should use a constant gain adaptive algorithm (as in Kouritzin [23]) to start and then switch to the type considered herein once close to steady state.

3.3 Theoretical Result

In this section, we provide our results.

3.3.1 Main Results

We will prove our results in a completely deterministic manner and then apply these results to each path. Therefore, we assume that $\chi \in (0, 1)$, d is a positive integer, $\{\bar{A}_k\}_{k=1}^{\infty}$ is a symmetric, positive semidefinite $R^{d\times d}$ -valued sequence, $\{\bar{b}_k\}_{k=1}^{\infty}$ is a $\mathbb{R}^d$-valued sequence and $\{\bar{h}_k\}_{k=1}^{\infty}$ is a $\mathbb{R}^d$-valued sequence satisfying:

$$\bar{h}_{k+1} = \bar{h}_k + \frac{1}{k^{\chi}}(\bar{b}_k - \bar{A}_k\bar{h}_k) \quad \text{for all} \quad k = 1, 2, ... \tag{3.13}$$

Our first main result establishes rates of convergence:

Theorem 3.1 *Let $\gamma \in [0, \chi)$, $h \in \mathbb{R}^d$ and A be symmetric and positive-definite.*

a) *If*

$$\lim_{n\to\infty}\left\|\frac{1}{n^{\chi}}\sum_{k=1}^{n}(\bar{A}_k - A)\right\| = 0, \quad \text{and} \quad \lim_{n\to\infty}\left|\frac{1}{n^{\chi-\gamma}}\sum_{k=1}^{n}(\bar{b}_k - \bar{A}_k h)\right| = 0, \tag{3.14}$$

then $n^{\gamma}|\bar{h}_n - h| \to 0$ a.s. as $n \to \infty$.

b) *Conversely,* $\lim_{n\to\infty}\left|\frac{1}{n^{\chi}}\sum_{k=1}^{n}(\bar{b}_k - \bar{A}_k h)\right| = 0$, *if* $\lim_{k\to\infty}\left|k^{1-\chi}(\bar{h}_k - h)\right| = 0$ *and*

$$\frac{1}{n^{\chi}}\sum_{k=1}^{n}k^{\chi-1}\left\|\bar{A}_k\right\| \quad \textit{is bounded in } n. \tag{3.15}$$

Now we state the almost sure version of the above theorem as the following corollary:

Corollary 3.1 *Let $\gamma \in [0, \chi)$, $h \in \mathbb{R}^d$ and A be symmetric and positive-definite.*

a) *If*

$$\lim_{n\to\infty}\left\|\frac{1}{n^{\chi}}\sum_{k=1}^{n}(A_k - A)\right\| = 0 \text{ and } \lim_{n\to\infty}\left|\frac{1}{n^{\chi-\gamma}}\sum_{k=1}^{n}(b_k - A_k h)\right| = 0 \quad a.s., \tag{3.16}$$

then $n^\gamma |h_n - h| \to 0$ *a.s. as* $n \to \infty$.

b) *Conversely,* $\lim\limits_{n\to\infty} \left| \frac{1}{n^\chi} \sum\limits_{k=1}^{n} (b_k - A_k h) \right| = 0\, a.s.$, *if* $\lim\limits_{k\to\infty} |k^{1-\chi}(h_k - h)| = 0\, a.s.$ *and* $\frac{1}{n^\chi} \sum\limits_{k=1}^{n} k^{\chi-1} \|A_k\|$ *is bounded in* n *almost surely.*

Proof. a) Fix ω such that (3.16) is true, recall (3.6); set $\bar{A}_k = A_k(\omega)$, $\bar{b}_k = b_k(\omega)$ and $\bar{h}_k = h_k(\omega)$ for all k; and apply Theorem 3.1. **b)** is similar. □

Corollary 3.1 implies $h_n(\omega)$, the solution of (3.2), converges to $h = A^{-1}b$ a.s.

Remark 3.1 *Lemma 3.1 of Appendix establishes that the first equation of (3.14) implies (3.15).*

Indeed, to establish the rate of convergence $n^\gamma |\bar{h}_n - h| \to 0$ a.s., one need only check standard conditions for the MSLLN in (3.16), which is less onerous task than checking the technical conditions in Corollary 1 or Corollary 3 in [34] say. Indeed, there appears to be a need for some extra stability in [34] by the imposition that "the real parts of the eigenvalues of A should be strictly less than a certain negative value depending on the asymptotic properties of $\{\gamma_n\}$ and $\{\delta_n\}$". We do not need any such extra condition.

Generally, we do not know h when using stochastic approximation so we cannot just verify second condition in (3.14) (or (3.16)) but rather use the following corollary instead of Theorem 3.1 (or Corollary 3.1).

Corollary 3.2 *Suppose* $\gamma \in [0, \chi)$ *and* A *is a symmetric positive-definite matrix.*

$$\frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (\bar{b}_k - b) \to 0 \text{ and } \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (\bar{A}_k - A) \to 0 \text{ a.s.} \qquad (3.17)$$

Then, $n^\gamma|\bar{h}_n - h| \to 0$ a.s. as $n \to \infty$.

Finally, we give a version of the theorem for linear processes under very general and verifiable conditions.

Theorem 3.2 *Let $\{\Xi_l\}$ be i.i.d. zero-mean random $\mathbb{R}^m$-vectors such that*

$$\sup_{t\geq 0} t^\alpha P(|\Xi_1|^2 > t) < \infty \quad \textit{for some } \alpha \in (1,2)$$

$(C_l)_{l\in\mathbb{Z}}$ be $\mathbb{R}^{(d+1)\times m}$-matrices such that $\sup_{l\in\mathbb{Z}} |l|^\sigma \|C_l\| < \infty$ for some $\sigma \in \left(\frac{1}{2}, 1\right]$,

$$(x_k^T, y_{k+1})^T = \sum_{l=-\infty}^{\infty} C_{k-l}\Xi_l,$$

$A_k = x_k x_k^T$, $b_k = y_{k+1}x_k$ and $A = E[x_k x_k^T]$ and $b = E[y_{k+1}x_k]$.
Then, $n^\gamma|h_n - h| \to 0$ a.s. as $n \to \infty$ a.s. for any $\gamma < \gamma_0(\chi) \doteq (\chi - \frac{1}{\alpha}) \wedge (\chi + 2\sigma - 2)$.

Remark 3.2 *Theorem 3.2 follows from Corollary 3.2 and Theorem 3.7 (to follow), by letting $\frac{1}{p} = \chi - \gamma$ and $\overline{X}_k^T = X_k^T = (x_k^T, y_{k+1})$ and correspondingly, $\overline{\Xi}_l = \Xi_l$, $\overline{C}_l = C_l$ and $\overline{\sigma} = \sigma$. σ and α are long-range dependence and heavy-tail parameters, respectively. Theorem 3.7 also appears in [20, Theorem 4].*

Remark 3.3 *$\sup_{l\in\mathbb{Z}} |l|^{\sigma''}\|C_l\| < \infty$ for some $\sigma'' > 1$ would be the (normal) short-range dependence and this clearly implies our weaker $\sup_{l\in\mathbb{Z}} |l|^\sigma\|C_l\| < \infty$ condition for some $\sigma \in \left(\frac{1}{2}, 1\right]$, which allows for long-range dependence. Our tail condition $\sup_{t\geq 0} t^\alpha P(|\Xi_1|^2 > t) < \infty$ is implied by the moment condition $E|\Xi_1|^{2\alpha} < \infty$. However, it too is general enough to allow non-standard (heavy-tailed) $\{A_k\}_{k=1}^\infty$, $\{b_k\}_{k=1}^\infty$, since $\alpha < 2$ corresponds to moments less than 2 for A_k and b_k.*

3.4 Marcinkiewicz Strong Law of Large Numbers for Partial Sums

Our basic assumptions are MSLLN for random variables for $\{A_k\}$ and $\{b_k\}$. (Technically, our assumptions are even more general as they allow the non-MSLLN case where $\chi - \gamma \leq \frac{1}{2}$ that could be verified by some other method in some special situations.) The beauty of this MSLLN assumption is that: 1) It is minimal in the sense that the linear algorithm with $A_k = I$ and $\mu_k = \frac{1}{k}$ reduces to the partial sums $h_{k+1} - h = \frac{1}{k}\sum\limits_{j=1}^{k}(b_j - b)$ (since $h = b$ when $A = I$) so a rate of convergence in the algorithm solution h_k implies a MSLLN for random variables $\{b_j\}$. 2) MSLLNs hold under very general conditions, including heavily-tailed and long-range dependent data. Hence, we review some of the literature in this area before giving simulation results for our theoretical work.

The classical independent case, due to Marcinkiewicz, is generalized slightly by Rio [29]:

Theorem 3.3 *Let $\{X_i\}$ be an m-dependent, identically distributed sequence of zero-mean $\mathbb{R}$-valued random variables such that $E|X_1|^p < \infty$ for some $p \in (1, 2)$. Then,*

$$\frac{1}{n^{\frac{1}{p}}}\sum_{i=1}^{n} X_i \to 0 \ a.s.$$

Actually, Rio gives a more general m-dependent result on page 922 of his work. However, the important observation for us is that only the p^{th} moment need be finite rather than a higher moment as is typical under some stronger dependence assumptions. Theorem 3.3 is quite useful in verifying our conditions when $\{A_k\}$ and $\{b_k\}$ may have heavily-tailed distributions but are independent or m-dependent. For example, if $\chi - \gamma \in \left(\frac{1}{2}, 1\right)$ and the $\{A_k\}$ and $\{b_k\}$ are defined as in (3.3) in terms of i.i.d. $\{x_k\}$ and $\{y_k\}$ with $E|x_1|^{\frac{2}{\chi-\gamma}} < \infty$, $E|y_1|^{\frac{2}{\chi-\gamma}} < \infty$, then $\{A_k\}_{k \geq M}$ and $\{b_k\}_{k \geq M}$ are identically distributed, M-dependent and

$$\frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (A_k - A) \to 0 \text{ and } \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{n} (b_k - b) \to 0$$

a.s., where $A = EA_k$ and $b = Eb_k$, by applying Theorem 3.3 for each component. Hence, (3.8) holds.

There are many other important results that include heavy-tails, long-range dependence or both. For example, Louhchi and Soulier [28] give the following result for linear symmetric α-stable (SαS) processes.

Theorem 3.4 *Let $\{\zeta_j\}_{j \in \mathbb{Z}}$ be i.i.d. sequence of SαS random variables with $1 < \alpha < 2$ and $\{c_j\}_{j \in \mathbb{Z}}$ be a bounded collection such that $\sum\limits_{j \in \mathbb{Z}} |c_j|^s < \infty$ for some $s \in [1, \alpha)$. Set $X_k = \sum\limits_{j \in \mathbb{Z}} c_{k-j} \zeta_j$. Then, for $p \in (1, 2)$ satisfying $\frac{1}{p} > 1 - \frac{1}{s} + \frac{1}{\alpha}$*

$$\frac{1}{n^{\frac{1}{p}}} \sum_{i=1}^{n} X_i \to 0 \textit{ a.s.}$$

The condition $s < \alpha$ ensures $\sum\limits_{j \in \mathbb{Z}} |c_j|^\alpha < \infty$ and thereby convergence of $\sum\limits_{j \in \mathbb{Z}} c_{k-j} \zeta_j$. Moreover, $\{X_k\}$ not only exhibits heavy tails but also long-range dependence if, for example, $c_j = |j|^{-\sigma}$ for $j \neq 0$ and some $\sigma \in \left(\frac{1}{2}, 1\right)$. Notice there is interactions between the heavy tail condition and the long-range dependent condition. In particular for a given p, heavier tails (α becomes smaller) implies that you cannot have as long-range dependence (s becomes smaller) and vice versa. Moreover, this result is difficult to apply in the stochastic approximation setting. For example, if wanted to apply it for $X_k = A_k$ in the scalar case, then we would need x_k such that $x_k^2 = A_k$ which is impossible when A_k is SαS.

One nice feature of mixing assumptions is that they usually transfer from random variables to functions (like squares) of random variables. There are many mixing results that handle long-range dependence. For example, Berbee [2] gives a nice β-mixing result. However, strong mixing is one of the most general types of mixing that is more easily verified in practice. Hence, we will quote the following strong mixing result from Rio [29] (Theorem 1) in terms of the inverse $\alpha^{-1}(u) = \sup\{t \in \mathbb{R}^+ : \alpha_{\lfloor t \rfloor} > u\}$ of the strong mixing coefficients

$$\alpha_n = \sup_{k \in \mathbb{Z}} \sup_{A \in \sigma(X_i, i \leq k-n), B \in \sigma(X_k)} |P(AB) - P(A)P(B)|$$

and the complementary quantile function

$$Q_X(u) = \sup\{t \in \mathbb{R}^+ : P(|X| > t) > u\}.$$

Theorem 3.5 *Let $\{X_i\}$ be an identically-distributed zero-mean sequence of $\mathbb{R}$-valued random variables such that $\int_0^1 [\alpha^{-1}(t/2)]^{p-1} Q_X^p(t)dt < \infty$ for some $p \in (1,2)$. Then,*

$$\frac{1}{n^{\frac{1}{p}}} \sum_{i=1}^{n} X_i \to 0 \ a.s.$$

Notice again that for a given p, heavier tails implies that you cannot have as long-range dependence and vice versa: If you wanted to maintain the same value of the integral condition and there became more area under $P(|X| > t)$, then there would be more area under $Q_X^p(t)$ so the area under $[\alpha^{-1}(t/2)]^{p-1}$, which is equal to $2 \sum\limits_{n=0}^{\infty} \alpha_n^{p-1}$, would have to decrease to compensate. Also, there can be difficulty in establishing that a given model satisfies the strong mixing condition with the required decay of mixing coefficients. Still, this is an important result for verifying our basic assumptions.

Another result in mixing area is given by Thanh, Yin and Wang [35, Theorem 3.11], whom considered MSLLN for double indexed and randomly weighted sums of mixing processes. Generally, they considered ρ^*-mixing types, which is defined as follows.

On the probability space $(\Omega, \mathcal{F}, P)$, let $\mathcal{A}$ and $\mathcal{B}$ be two sub-σ-algebras of $\mathcal{F}$. We denote by $\mathcal{L}_2(\mathcal{A})$ the space of all square integrable and $\mathcal{A}$-measurable random variables. The maximal coefficient of correlation is defined by

$$\rho(\mathcal{A}, \mathcal{B}) = \sup_{f \in \mathcal{L}_2(\mathcal{A}), g \in \mathcal{L}_2(\mathcal{B})} |corr(f, g)|.$$

Let $\{X_n, n \geq 1\}$ be a sequence of random variables. For a subset S of $\mathbb{N} = \{1, 2, ...\}$, $\sigma(S)$ means the σ-field generated by $\{X_n, n \in S\}$. For $n \geq 1$, define $\rho_n^* \doteq \rho^*(X, n) \doteq \sup \rho(\sigma(S), \sigma(T))$, where the supremum is taken over all pairs of nonempty finite sets S, T of $\mathbb{N}$ such that $dist(S, T) = \inf_{s \in S, t \in T} |s - t| \geq n$. The sequence $\{X_n, n \geq 1\}$ is said to be ρ^*-mixing if $\rho_n^* \to 0$ as $n \to \infty$.

Theorem 3.6 *Let $0 \leq r < 1$, and let N be a positive integer. Let $1 \leq p < 2$, and let $\{X_n, n \geq 1\}$ be a sequence of mean zero strictly stationary random variables such that $\rho^*(X, N) \leq r$. Suppose that $\{A_{ni}, n \geq 1, 1 \leq i \leq n\}$ is an array of random variables such that, for each $n \geq 1$, the sequence $A_n = \{A_{ni}, 1 \leq i \leq n\}$ satisfies $\rho^*(A_n, N) \leq r$, and*

$$\sum_{i=1}^{n} E(|A_{ni}|^q) = O(n) \qquad \textit{for some} \quad q > \frac{2p}{2-p}. \tag{3.18}$$

If $E|X_1|^{2p} < \infty$, and $\{A_{ni}, n \geq 1, 1 \leq i \leq n\}$ is independent of $\{X_i, i \geq 1\}$, *then*

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{i=1}^{n} A_{ni} X_i = 0 \quad a.s. \tag{3.19}$$

Notice that the above Theorem looked at the the moments of A_{ni} and X_i separately. However, if we look at the moment of products, we find by Hölder's inequality that $E|A_{ni}X_i|^{\frac{2p}{3-p}} < \infty$ and $\frac{2p}{3-p} \in [1, 4)$, so some heavy-tail situations are allowed. However, the ρ^*-mixing condition does not allow long-range dependence situations, since

$$\sum_{k=1}^{\infty} E[X_1 X_k] = \sum_{j=0}^{\infty} \sum_{k=1}^{N} E[X_1 X_{jN+k}] \leq N \sum_{j=0}^{\infty} r^j < \infty.$$

A new MSLLN for outer products of multivariate linear processes with long-range dependence and heavy tails is studied in [20]. A new decoupling property is proved that shows the convergence rate is determined by the worst of the heavy tails or the long-range dependence, but not the combination. This result used to obtain Marcinkiewicz Strong Law of Large Numbers for stochastic approximation (Theorem 3.2). The result is as follows.

Theorem 3.7 *Let $\{\Xi_l\}$ and $\{\overline{\Xi}_l\}$ be i.i.d. zero mean random $\mathbb{R}^m$-vectors such that $\Xi_l = \left(\xi_l^{(1)}, ..., \xi_l^{(m)}\right)$, $\overline{\Xi}_l = \left(\overline{\xi}_l^{(1)}, ..., \overline{\xi}_l^{(m)}\right)$, $E[|\Xi_1|^2] < \infty$, $E[|\overline{\Xi}_1|^2] < \infty$ and $\max\limits_{1\leq i,j\leq m} \sup\limits_{t\geq 0} t^{\alpha} P(|\xi_1^{(i)} \overline{\xi}_1^{(j)}| > t) < \infty$ for some $\alpha \in (1, 2)$. Moreover, suppose matrix sequences $(C_l)_{l\in\mathbb{Z}}, (\overline{C}_l)_{l\in\mathbb{Z}} \in \mathbb{R}^{(d+1)\times m}$ satisfy*

$$\sup_{l\in\mathbb{Z}} |l|^{\sigma} \|C_l\| < \infty, \sup_{l\in\mathbb{Z}} |l|^{\overline{\sigma}} \|\overline{C}_l\| < \infty \quad \textit{for some} \quad (\sigma, \overline{\sigma}) \in \left(\frac{1}{2}, 1\right],$$

X_k, $\overline{X}_k$ take form of (3.4), $D_k = X_k \overline{X}_k^T$ and $D = E[X_1 \overline{X}_1^T]$. Then, for p

satisfying $p < \frac{1}{2-\sigma-\overline{\sigma}} \wedge \alpha$

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (D_k - D) = 0 \qquad a.s.$$

Remark 3.4 *This theorem actually shows the MSLLN for* $D_k - E[D_k]$*, where* $D_k = \begin{pmatrix} x_k x_k^T & y_{k+1} x_k \\ y_{k+1} x_k^T & y_{k+1}^2 \end{pmatrix}$*, which is more than required, so we can throw out the unneeded columns.*

3.5 Experimental Results

In this section we now verify our results of the previous section experimentally in the stochastic approximation setting discussed in the introduction. In particular, we use *power law* or *folded t* distributions.

Power law distribution: A random variable ξ obeys a power law with parameters $\beta > 1$ and $x_{\min} > 0$, written $\xi \sim PL(x_{min}, \beta)$, if it has density

$$f(x) = \frac{\beta - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\beta} \qquad \forall\ x \geq x_{min}$$

Note that $E|\xi|^r = \begin{cases} x_{min}^r (\frac{\beta-1}{\beta-1-r}) & r < \beta - 1 \\ \infty & r \geq \beta - 1 \end{cases}$.

Folded t distribution: A non-negative random variable ξ has a folded t distribution with parameter $\beta > 1$, written $\xi \sim Ft(\beta)$, if it has density

$$f(x) = \frac{2\Gamma(\frac{\beta}{2})}{\Gamma(\frac{\beta-1}{2})\sqrt{(\beta-1)\pi}} \left(1 + \frac{x^2}{\beta - 1}\right)^{-\frac{\beta}{2}} \qquad \forall\ x > 0.$$

Note that $E(|\xi|^r)$ exists if and only if $r < \beta - 1$.

Experimental results in this section are divided in two parts.

3.5.1 Heavy-Tailed Cases

Assume $N = 1$ in (3.3), dimension is $d = 2$ and $\{(x_k^{(1)}, x_k^{(2)}, \epsilon_k)^T,\ k = 1, 2, ...\}$ are i.i.d. random vectors so linear algorithm (3.2) reduces to:

$$h_{k+1} = h_k + \mu_k(x_k y_{k+1} - x_k x_k^T h_k) = h_k + \mu_k(x_k x_k^T h + x_k \epsilon_k - x_k x_k^T h_k). \quad (3.20)$$

For consistency and performance, we always let $x_k^{(1)}, x_k^{(2)}$ and ϵ_k be independent. The runs are always initialized with $h_1 = (101, 101)^T$ and, for testing purposes, the optimal $h = (1, 1)^T$ is known.

Example 3.1 *Let* $x_k^{(1)}, x_k^{(2)} \sim PL(x_{min} = 1, \beta)$ *and* $\epsilon_k = \epsilon_k' - E(\epsilon_k')$ *with* $\epsilon_k' \sim PL(x_{min}' = 0.01, \beta)$*. The normalized errors in 100 trial simulations,* $\{h_n^{(i)}\}_{i=1}^{100}$*, are averaged* $\overline{rh} = \frac{1}{100}\sum_{i=1}^{100} \frac{|h_n^{(i)} - h|}{|h_1 - h|}$ *and given in the Table 3.1 in terms of gain parameter* χ*, distributional parameter* β *and sample size* n*.*

Table 3.1: Algorithm performance-Power Law

	n=100000			n=750000			n=1500000		
$\chi\backslash\beta$	3.5	4	4.5	3.5	4	4.5	3.5	4	4.5
0.6	0.0864	0.0314	0.0169	0.0707	0.0243	0.0115	0.0548	0.0203	0.0099
0.7	0.0525	0.0190	0.0098	0.0487	0.0159	0.0067	0.0457	0.0141	0.0056
0.75	0.0397	0.0151	0.0082	0.0449	0.0137	0.0051	0.0456	0.0114	0.0042
0.8	0.0326	0.0136	0.0105	0.0448	0.0111	0.0038	0.0402	0.0087	0.0031
0.85	0.0314	0.0168	0.0549	0.0398	0.0085	0.0082	0.0324	0.0070	0.0035
0.9	0.0344	0.0719	0.2445	0.0438	0.0118	0.0764	0.0272	0.0079	0.0341
0.95	0.0902	0.3047	0.6631	0.3739	0.0897	0.3068	0.0248	0.0519	0.1963
0.98	0.2226	0.5733	1.0154	0.9219	0.2302	0.5251	0.0374	0.1488	0.3930
1	0.3876	0.8062	0.6631	1.1891	0.3745	0.6925	0.0662	0.2596	0.5644

The Marcinkiewicz threshold, $M = \frac{2}{(\beta-1)}$*, corresponding to* $\beta = 3.5$*,* $\beta = 4$ *and* $\beta = 4.5$ *are respectively* $M = 0.8$*,* 0.67 *and* 0.57*. Our theoretical results prove convergence above this threshold. While the results in Table 3.1* *are obviously still influenced by (heavy-tailed) randomness, one can see that convergence does appear to be taking place as one moves from $n = 100,000$ through $n = 750,000$ to $n = 1,500,000$ when $\chi > M$ and it is less clear that convergence is taking place when $\chi < M$. Furthermore, our (as well as prior) theoretical results predict rates of convergence that increase in χ. Indeed, in the case $\beta = 4$ our theoretical results suggest that $\chi \approx 1$ should result in a rate of convergence $n^{0.33}|h_n - h| \to 0$ a.s. while $\chi = 0.85$ should only result in a rate of convergence $n^{0.18}|h_n - h| \to 0$ a.s. Conversely, Table 3.1 demonstrates that $\chi = 0.85$ performs better, which seems to contradict the theory. However, this paradox is explained by the exploding constants discussion of the introduction and, in fact, points out that more refined theory, involving functional results, is needed. The proper way to use our theoretical results then is to predict the best χ (lowest value of $\overline{rh}$) in the range of $(M, 1]$ i.e. in $(0.8, 1]$, $(0.67, 1]$ and $(0.57, 1]$, respectively for our three β's.*

Table 3.2: Best fixed χ-Power Law

	n=100000			n=750000			n=1500000		
β	3.5	4	4.5	3.5	4	4.5	3.5	4	4.5
Best χ	0.85	0.8	0.75	0.85	0.85	0.8	0.95	0.85	0.8
Resulting γ	0.05	0.13	0.18	0.05	0.18	0.23	0.15	0.18	0.23

The best χ's, corresponding to the smallest value of $\overline{rh}$ for $\beta = 3.5$, $\beta = 4$ and $\beta = 4.5$ and 3 different sample sizes, as well as the γ corresponding to the theoretical rate of convergence $o(n^{-\gamma})$ are summarized in Table 3.2. In all cases, the best value for χ is in the predicted range. As we explained, a faster decreasing gain is appropriate for a heavier-tailed distribution, which is also confirmed by Table 3.2. Notice also that the best χ increases in n, a phenomenon consistent with our exploding constants and the initial condition effect discussion, which suggests that we might do better by letting χ increase *in k, perhaps starting at M and heading towards* 1.

Table 3.3: Increasing χ comparison-Power Law

β	3.5	4	4.5
Best χ	0.0248	0.0070	0.0031
Increasing χ	0.0240	0.0068	0.0030

Table 3.3 compares values of $\overline{rh}$ between our best fixed and increasing $\chi = \chi_k$ when the sample size is $n = 1,500,000$. For clarity, our gain at step k is now $\mu_k = \mu_0 k^{-\chi_k}$. In this increasing case, the initial χ is (β-dependent) Marcinkiewicz threshold (when $k = 0$) and χ_k increases as $k \to 1,500,000$. For simplicity, we just took $\chi_k = \tanh^{-1}(a+bk+ck^2)$ and estimated a, b and c via least squares with the data points $(k = 0; \chi = M)$, $(k = 100,000; \chi = 0.8)$, $(k = 750,000; \alpha = 0.85)$ and $(k = 1,500,000; \alpha = 0.85)$ in the $\beta = 4$ case. (In the other cases, we followed the same plan using the best χ for a given n.) The result show improved performance over the best constant χ.

Now, we repeat the previous example with a different distribution. Since the results are consistent with those of the previous example, we will keep our discussion to a minimum.

Example 3.2 *Let $x_k^{(1)}, x_k^{(2)} \sim Ft(\beta)$ and $\epsilon_k = \epsilon'_k - E(\epsilon'_k)$ with $\epsilon'_k \sim Ft(\beta)$. The simulation results for three β's: $3.5, 4$ and 4.5 with corresponding Marcinkiewicz thresholds, $M = \frac{2}{\beta-1}$, 0.8, 0.67 and 0.57 are given in Table 3.4 with sample sizes: $n = 50,000, 100,000$ and $750,000$.*

Table 3.4: Algorithm performance-Folded t

	n=50000			n=100000			n=750000		
$\chi\backslash\beta$	3.5	4	4.5	3.5	4	4.5	3.5	4	4.5
0.6	0.0958	0.0345	0.0221	0.0929	0.0336	0.0177	0.0590	0.0195	0.0104
0.7	0.0697	0.0245	0.0138	0.0661	0.0216	0.0112	0.0318	0.0120	0.0064
0.75	0.0599	0.0204	0.0113	0.0556	0.0173	0.0089	0.0336	0.0099	0.0050
0.8	0.0505	0.0172	0.0103	0.0439	0.0140	0.0075	0.0374	0.0076	0.0038
0.85	0.0399	0.0145	0.0098	0.0341	0.0118	0.0063	0.0339	0.0058	0.0029
0.9	0.0312	0.0133	0.0087	0.0278	0.0100	0.0057	0.0265	0.0048	0.0024
0.95	0.0275	0.0241	0.0097	0.0245	0.0089	0.0060	0.0205	0.0039	0.0021
0.98	0.0347	0.0475	0.0212	0.0274	0.0117	0.0121	0.0179	0.00371	0.0032
0.99	0.0404	0.0583	0.0295	0.0310	0.0149	0.0172	0.0173	0.00373	0.0048
1	0.0486	0.0700	0.0413	0.0369	0.0205	0.0249	0.0170	0.0039	0.0077

A summary of of best χ result is given in Table 3.5. Again, a smaller β corresponds to heavier tails and larger best χ. Moreover, as we predicted the best χ for $\beta = 3.5$, $\beta = 4$ and $\beta = 4.5$ in the range of $(0.8, 1]$, $(0.67, 1]$ and $(0.57, 1]$, respectively. Best χ's increase in sample size.

Table 3.5: Best fixed χ-Folded t

	n=50000			n=100000			n=750000		
β	3.5	4	4.5	3.5	4	4.5	3.5	4	4.5
Best χ	0.95	0.9	0.9	0.95	0.95	0.9	1	0.98	0.95
$\gamma <$	0.15	0.23	0.33	0.15	0.28	0.33	0.2	0.31	0.38

Table 3.6 shows the performance of algorithm improves by going to an increasing χ when $n = 750,000$. Again, χ starts at Marcinkiewicz threshold

Table 3.6: Increasing χ comparison-Folded t

β	3.5	4	4.5
Best χ	0.0170	0.00371	0.0021
Increasing χ	0.0166	0.00364	0.0019

for each choice of β *and increases a function of sample size* $1 \leq n \leq 750,000$. *The function is* $\chi = \tanh^{-1}(a' + b'n + c'n^2)$ *and coefficients* a', b' *and* c' *are estimated with least square method.*

3.5.2 Combined Heavy-Tailed and Long-Range Dependence Case

If we take $N = 1$ and dimension $d = 1$, we have $(x_k, y_{k+1}) = \sum\limits_{j=-\infty}^{\infty} C_{k-j}\Xi_j$, in which $C_j = (c_j, c_j)^T$ and $\Xi_j = (\xi_j^{(1)}, \xi_j^{(2)})_{j\in\mathbb{Z}}$ are i.i.d.. Hence, $x_k = \sum\limits_{j\in\mathbb{Z}} c_{k-j}\xi_j^{(1)}$ and $y_{k+1} = \sum\limits_{j\in\mathbb{Z}} c_{k-j}\xi_j^{(2)}$, where $\xi_j^{(2)} = h\xi_j^{(1)} + a_j$ and $\{a_j\}$'s are i.i.d. zero mean random variables. This relation between $\xi_j^{(1)}$ and $\xi_j^{(2)}$ is due to the fact that $y_{k+1} = x_k h + \epsilon_k$ and $\epsilon_k = \sum\limits_{j\in\mathbb{Z}} c_{k-j}a_j$. We consider $\{c_j = |j|^{-\sigma}\}, \text{for}\, j \neq 0$ and $\sigma \in (\frac{1}{2}, 1]$, $c_0 = 1$. The linear algorithm (3.2) reduces to:

$$h_{k+1} = h_k + \mu_k(x_k y_{k+1} - x_k^2 h_k) = h_k + \mu_k(x_k^2 h + x_k\epsilon_k - x_k^2 h_k). \quad (3.21)$$

The initial and optimal values are $h_1 = 401$ and $h = 1$.

Example 3.3 *Let* $\xi_j^{(1)} \sim PL(x_{min} = 0.01, \beta)$ *and* $a_j = f_j - E(f_j)$ *with* $f_j \sim PL(x'_{min} = 0.01, \beta)$. *The simulation is done for one-sided process and since in computer we cannot technically do infinite sum, we assume summation over the range of* $(0, 500, 000)$. *Similarly, the normalized errors in 100 trial simulations,* $\{h_n^{(i)}\}_{i=1}^{100}$, *are averaged and results for different* χ*'s,* β*'s and* *sample sizes n are presented in the following tables. The assumed σ is* 0.65*. The Marcinkiewicz threshold, $M = \frac{1}{\alpha} \vee (2 - 2\sigma)$, corresponding to $\beta = 4$, $\beta = 4.5$ and $\beta = 5$ is* 0.7*. Hence, predicted ranges for χ's with smallest $\overline{rh}$ will be* $(0.7, 1]$*. Simulation results are provided in Table 3.7 with summary of best χ in Table 3.8. It worth noticing that the convergence does not seem to take place below the Marcinkiewicz threshold and the best χs are in the predicted ranges and the normalized error decreases as β increases.*

Note that by considering $\sigma = 0.65$, the maximum of $2 - 2\sigma$ and $\frac{1}{\alpha}$ for all $\beta = 4, 4.5$ and 5 is $2 - 2\sigma$, hence we do not expect much change in the χ as β changes. In addition, the rate of convergence for all considered β's is determined by $\gamma < \chi + 2\sigma - 2$.

Table 3.7: Algorithm performance for LRD-HT cases with $\sigma = 0.65$

	n=100			n=5000			n=10,000		
$\chi \backslash \beta$	4	4.5	5	4	4.5	5	4	4.5	5
0.6	0.010917	0.006166	0.004508	0.013172	0.007826	0.005897	0.012465	0.007359	0.005527
0.7	0.000665	0.000237	0.000132	0.000958	0.000414	0.000262	0.000881	0.000377	0.000238
0.75	2.98e-05	7.88e-06	6.49e-06	9.77e-05	3.15e-05	1.70e-05	8.79e-05	2.83e-05	1.52e-05
0.8	1.02e-05	7.76e-06	6.39e-06	5.19e-06	3.80e-06	3.11e-06	4.72e-06	3.30e-06	2.69e-06
0.85	9.91e-06	7.77e-06	6.41e-06	5.01e-06	3.91e-06	3.21e-06	4.38e-06	3.37e-06	2.76e-06
0.9	9.93e-06	7.79e-06	6.45e-06	5.20e-06	4.12e-06	3.39e-06	4.54e-06	3.50e-06	2.86e-06
0.95	1.00e-05	7.90e-06	6.55e-06	5.63e-06	4.42e-06	3.62e-06	4.73e-06	3.65e-06	2.98e-06
0.98	1.01e-05	7.99e-06	6.61e-06	5.97e-06	4.69e-06	3.86e-06	4.88e-06	3.77e-06	3.08e-06
1	1.02e-05	8.04e-06	6.65e-06	6.28e-06	4.91e-06	4.03e-06	5.02e-06	3.89e-06	3.19e-06

Table 3.8: Best fixed χ-Power Law, LRD with $\sigma = 0.65$

	n=100			n=5000			n=10000		
β	4	4.5	5	4	4.5	5	4	4.5	5
Best χ	0.85	0.8	0.8	0.85	0.8	0.8	0.85	0.8	0.8
Resulting γ	0.15	0.1	0.1	0.15	0.1	0.1	0.15	0.1	0.1

3.6 The Proof of Theorem 3.1

Part a) Step 1: Reduce rate of convergence to convergence of a transformed algorithm.

Letting $\eta_k = \left(\frac{k+1}{k}\right)^\gamma - 1$, setting $g_k = k^\gamma \left(\bar{h}_k - h\right)$ and using (3.13), one finds that

$$g_{k+1} \quad = \quad g_k + \frac{1}{k^\chi}\left(\hat{b}_k - \hat{A}_k g_k\right) + \eta_k g_k, \tag{3.22}$$

where

$$\hat{b}_k = (k+1)^\gamma \left(\bar{b}_k - \bar{A}_k h\right) \text{ and } \hat{A}_k = \left(\frac{k+1}{k}\right)^\gamma \bar{A}_k. \tag{3.23}$$

However, we have by Taylor's theorem and assumption that

$$\frac{1}{n^\chi}\sum_{k=1}^{n} \eta_k \|A\| \quad \leq \quad \frac{\gamma}{n^\chi}\sum_{k=1}^{n} k^{-1} \to 0, \quad \text{as } n \to \infty. \tag{3.24}$$

Step 2: Show MSLLN for new coefficients i.e. $\frac{1}{n^\chi}\sum\limits_{k=1}^{n}(\hat{A}_k - A) \to 0$, and $\frac{1}{n^\chi}\sum\limits_{k=1}^{n}\hat{b}_k \to 0 \quad \text{as } n \to \infty$.

$$
\begin{aligned}
& \left\| \frac{1}{n^\chi}\sum_{k=1}^{n}\left(\frac{k+1}{k}\right)^\gamma (\bar{A}_k - A) - \frac{2^\gamma}{n^\chi}\sum_{k=1}^{n}(\bar{A}_k - A) \right\| \\
= \; & \left\| \frac{1}{n^\chi}\sum_{k=2}^{n}\sum_{j=2}^{k}\left[\left(\frac{j+1}{j}\right)^\gamma - \left(\frac{j}{j-1}\right)^\gamma\right](\bar{A}_k - A) \right\| \\
\le \; & \sum_{j=2}^{n}\left[\left(\frac{j}{j-1}\right)^\gamma - \left(\frac{j+1}{j}\right)^\gamma\right]\frac{1}{n^\chi}\left(\left\|\sum_{k=2}^{n}(\bar{A}_k - A)\right\| + \left\|\sum_{k=2}^{j-1}(\bar{A}_k - A)\right\|\right) \\
\le \; & \frac{1}{n^\chi}\left\|\sum_{k=2}^{n}(\bar{A}_k - A)\right\|\left[2^\gamma - \left(\frac{n+1}{n}\right)^\gamma\right] \\
+ \; & \sum_{j=2}^{n}\left[\left(\frac{j}{j-1}\right)^\gamma - \left(\frac{j+1}{j}\right)^\gamma\right]\left(\frac{j-2}{n}\right)^\chi \frac{1}{(j-2)^\chi}\left\|\sum_{k=2}^{j-1}(\bar{A}_k - A)\right\|
\end{aligned}
$$

which goes to zero by assumption and the Toeplitz lemma. By Taylor's theorem

$$
\begin{aligned}
& \left| \frac{1}{n^\chi}\sum_{k=1}^{n}(k+1)^\gamma\left(\bar{b}_k - \bar{A}_k h\right) - \frac{1}{n^\chi(n+1)^{-\gamma}}\sum_{k=1}^{n}\left(\bar{b}_k - \bar{A}_k h\right) \right| \\
= \; & \frac{1}{n^\chi}\left|\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}\left[j^\gamma - (j+1)^\gamma\right]\left(\bar{b}_k - \bar{A}_k h\right)\right| \\
\le \; & \frac{1}{n^\chi}\sum_{j=2}^{n}\gamma j^{\gamma-1}\left|\sum_{k=1}^{j-1}\left(\bar{b}_k - \bar{A}_k h\right)\right| \\
\le \; & \frac{\gamma}{n^\chi}\sum_{j=2}^{n} j^{\chi-1}\frac{1}{(j-1)^{\chi-\gamma}}\left|\sum_{k=1}^{j-1}\left(\bar{b}_k - \bar{A}_k h\right)\right|,
\end{aligned}
\tag{3.25}
$$

which goes to zero by the Toeplitz lemma.

Step 3: Convergence of g_k, hence the rate of convergence of $\bar{h}_k$ follows from the Proposition 3.1 with $b = 0$, $\hat{h}_k = g_k$, $h = 0$ and $\eta_k = \left(\frac{k+1}{k}\right)^\gamma - 1$. □

Proposition 3.1 *Suppose $\{\hat{A}_k\}_{k=1}^{\infty}$ is a symmetric, positive-semidefinite $R^{d\times d}$-valued sequence; A is a (symmetric) positive-definite matrix; $\chi \in (0,1)$; $\theta \in (\chi, 1]$; $\eta_k \le \frac{\bar{\eta}}{k^\theta}$; $\bar{\eta} > 0$ and*

$$\hat{h}_{k+1} = \hat{h}_k + \frac{1}{k^\chi}(\hat{b}_k - \hat{A}_k\hat{h}_k) + \eta_k\hat{h}_k \quad \textit{for all} \quad k = 1, 2, ...; \tag{3.26}$$

$$\frac{1}{n^\chi}\sum_{k=1}^{n}(\hat{b}_k - b) \to 0 \quad \textit{and} \quad \frac{1}{n^\chi}\sum_{k=1}^{n}(\hat{A}_k - A) \to 0. \tag{3.27}$$

Then, $\hat{h}_n \to h \doteq A^{-1}b$ as $n \to \infty$.

Notation: To ease the notation in the sequel, we will take the product over no factors to be 1 and the sum of no terms to be 0. For convenience, we let:

$$\nu_k := \hat{h}_k - h, \qquad Y_k := \hat{A}_k - A, \qquad z_k := \hat{b}_k - \hat{A}_k h. \tag{3.28}$$

Proof. Step 1: Show simplified algorithm with A_k's replaced converges. We note $\dfrac{1}{n^\chi}\displaystyle\sum_{k=1}^{n} z_k \to 0$ and will show $\nu_k \to 0$, by proving $u_k \to 0$ and $w_k := \nu_k - u_k \to 0$, where

$$u_{k+1} = \left(I - \frac{A}{k^\chi} + \eta_k I\right)u_k + \frac{z_k}{k^\chi} + \eta_k h \quad \text{subject to} \quad u_1 = \nu_1. \tag{3.29}$$

By induction, we have:

$$u_n = \prod_{l=1}^{n-1}\left(I - \frac{A}{l^\chi} + \eta_l I\right)u_1 + \sum_{j=1}^{n-1} F_{j,n} z_j + \sum_{j=1}^{n-1} \bar{F}_{j,n} h \quad \text{for } n = 1, 2, ... \tag{3.30}$$

where

$$\begin{cases} F_{j,n} = \frac{1}{j^{\chi}} \prod_{l=j+1}^{n-1} \left(I - \frac{A}{l^{\chi}} + \eta_l I \right) \\ \bar{F}_{j,n} = \eta_j j^{\chi} F_{j,n} \, \mathrm{for} j = 1, 2, ..., n-1, n = 2, 3, ... \end{cases} \tag{3.31}$$

Hence,by (3.30), (3.31) and Lemma 3.2 i, ii)

$$\begin{aligned} \lim_{n\to\infty} |u_n| \quad &\leq \quad \lim_{n\to\infty} \left\| \prod_{l=1}^{n-1} \left(I - \frac{A}{l^{\chi}} + \eta_l I \right) \right\| |u_1| \\ &+ \quad \lim_{n\to\infty} \left| \sum_{j=1}^{n-1} F_{j,n} z_j \right| + \lim_{n\to\infty} \left| \sum_{j=1}^{n-1} \bar{F}_{j,n} h \right| = 0. \end{aligned} \tag{3.32}$$

Step 2: Transfer stability from A to blocks of A_k.

Define the blocks

$$\begin{cases} n_k = \lfloor (ak)^{\frac{1}{1-\chi}} \rfloor := \max\{ i \in N_0 : i \leq (ak)^{\frac{1}{1-\chi}} \} \\ I_k = \{ n_k, n_k + 1, \cdots, n_{k+1} - 1 \} \end{cases} \tag{3.33}$$

for $k = 0, 1, 2, ...$ and the block products

$$U_k = \prod_{l \in I_k} \left(I - \frac{\hat{A}_l}{l^{\chi}} + \eta_l I \right) \, \mathrm{and} \, V_{j,k} = \prod_{l=j+1}^{n_{k+1}-1} \left(I - \frac{\hat{A}_l}{l^{\chi}} + \eta_l I \right) \frac{1}{j^{\chi}} Y_j. \tag{3.34}$$

For the U_k's we have

$$\begin{aligned} \prod_{l \in I_k} \left(I - \frac{\hat{A}_l}{l^{\chi}} + \eta_l I \right) = I - \sum_{l \in I_k} \frac{\hat{A}_l}{l^{\chi}} + \sum_{l \in I_k} \eta_l I + \sum_{\substack{l_1, l_2 \in I_k \\ l_1 > l_2}} \left(\frac{\hat{A}_{l_1}}{l_1^{\chi}} - \eta_{l_1} I \right) \left(\frac{\hat{A}_{l_2}}{l_2^{\chi}} - \eta_{l_2} I \right) \\ - \sum_{\substack{l_1, l_2, l_3 \in I_k \\ l_1 > l_2 > l_3}} \left(\frac{\hat{A}_{l_1}}{l_1^{\chi}} - \eta_{l_1} I \right) \left(\frac{\hat{A}_{l_2}}{l_2^{\chi}} - \eta_{l_2} I \right) \left(\frac{\hat{A}_{l_3}}{l_3^{\chi}} - \eta_{l_3} I \right) + \cdots (-1)^k \prod_{l \in I_k} \left(\frac{\hat{A}_l}{l^{\chi}} - \eta_l I \right) \end{aligned}$$

so

$$\begin{aligned}
\|U_k\| &\leq \left\| I - \sum_{l \in I_k} \frac{\hat{A}_l}{l^\chi} \right\| + \sum_{l \in I_k} \eta_l + \left\| \sum_{\substack{l_1, l_2 \in I_k \\ l_1 > l_2}} \left(\frac{\hat{A}_{l_1}}{l_1^\chi} - \eta_{l_1} I \right) \left(\frac{\hat{A}_{l_2}}{l_2^\chi} - \eta_{l_2} I \right) \right\| \\
&+ \left\| \sum_{\substack{l_1, l_2, l_3 \in I_k \\ l_1 > l_2 > l_3}} \left(\frac{\hat{A}_{l_1}}{l_1^\chi} - \eta_{l_1} I \right) \left(\frac{\hat{A}_{l_2}}{l_2^\chi} - \eta_{l_2} I \right) \left(\frac{\hat{A}_{l_3}}{l_3^\chi} - \eta_{l_3} I \right) \right\| \\
&+ \cdots + \prod_{l \in I_k} \left\| \frac{\hat{A}_l}{l^\chi} - \eta_l I \right\|. \qquad (3.35)
\end{aligned}$$

However, we know that $\sum_{j_1 > j_2 > \cdots > j_k} a_{j_1} a_{j_2} \cdots a_{j_k} \leq \frac{1}{k!} \left(\sum_j a_j \right)^k$ for $a_j \geq 0$ so, it follows that

$$\begin{aligned}
&\sum_{\substack{l_1, l_2 \in I_k \\ l_1 > l_2}} \left\| \frac{\hat{A}_{l_1}}{l_1^\chi} - \eta_{l_1} I \right\| \left\| \frac{\hat{A}_{l_2}}{l_2^\chi} - \eta_{l_2} I \right\| \\
&\quad + \sum_{\substack{l_1, l_2, l_3 \in I_k \\ l_1 > l_2 > l_3}} \left\| \frac{\hat{A}_{l_1}}{l_1^\chi} - \eta_{l_1} I \right\| \left\| \frac{\hat{A}_{l_2}}{l_2^\chi} - \eta_{l_2} I \right\| \left\| \frac{\hat{A}_{l_3}}{l_3^\chi} - \eta_{l_3} I \right\| + \cdots + \prod_{l \in I_k} \left\| \frac{\hat{A}_l}{l^\chi} - \eta_l I \right\| \\
&\quad \leq \sum_{m=2}^{n_{k+1} - n_k} \frac{\left(\sum_{l \in I_k} \left(\frac{\|\hat{A}_l\|}{l^\chi} + \eta_l \right) \right)^m}{m!}.
\end{aligned}$$

As a result, we find by (3.35) that

$$\begin{aligned}
\|U_k\| &\leq \left\| I - A \sum_{l \in I_k} \frac{1}{l^\chi} \right\| + \left\| \sum_{l \in I_k} \frac{Y_l}{l^\chi} \right\| + \sum_{l \in I_k} \eta_l \\
&+ \sum_{m=2}^{n_{k+1} - n_k} \frac{\left(\sum_{l \in I_k} \left(\frac{\|\hat{A}_l\|}{l^\chi} + \eta_l \right) \right)^m}{m!}. \qquad (3.36)
\end{aligned}$$

Now, let λ_{min} and λ_{max} be the smallest and biggest eigenvalues of A and define $a' = \frac{a}{1-\chi}$, where $a > 0$ is chosen small enough that

$$a' \leq \left\{ \frac{2}{\lambda_{min} + \|A\|}, \frac{1}{d\|A\|}, \frac{\lambda_{min}}{e^1 (d\|A\|)^2} \right\}. \tag{3.37}$$

Then, by (3.33) and the fact that

$$\frac{1}{1-\chi}(n_{k+1}^{1-\chi} - n_k^{1-\chi}) \leq \sum_{l \in I_k} \frac{1}{l^\chi} \leq \frac{1}{1-\chi}((n_{k+1}-1)^{1-\chi} - (n_k - 1)^{1-\chi})$$

we have $\lim_{k\to\infty} \left(\sum_{l\in I_k} \frac{1}{l^\chi} - a' \right)$ is in the range of

$$\left(\lim_{k\to\infty} \frac{n_{k+1}^{1-\chi} - n_k^{1-\chi} - a}{1-\chi} \;,\; \lim_{k\to\infty} \frac{n_{k+1}^{1-\chi} - n_k^{1-\chi} - a}{1-\chi} + \frac{n_k^{1-\chi} - (n_k - 1)^{1-\chi}}{1-\chi} \right),$$

so by Taylor's theorem

$$\begin{aligned} \lim_{k\to\infty} \left| \sum_{l\in I_k} \frac{1}{l^\chi} - a' \right| &\leq \lim_{k\to\infty} \left\{ \frac{1}{1-\chi} \left| n_{k+1}^{1-\chi} - n_k^{1-\chi} - a \right| + \frac{1}{(n_k - 1)^\chi} \right\} \\ &= 0, \end{aligned} \tag{3.38}$$

which also implies

$$\lim_{k\to\infty} \sum_{l\in I_k} \eta_l \leq \bar{\eta} \lim_{k\to\infty} n_k^{\chi-\theta} \sum_{l\in I_k} \frac{1}{l^\chi} = 0. \tag{3.39}$$

For arbitrary $\epsilon > 0$ one finds some $K_\epsilon > 0$ by (3.38) and (3.37) such that

$$\begin{aligned} \left\| I - A \sum_{l\in I_k} \frac{1}{l^\chi} \right\| &= \max \left\{ \|A\| \sum_{l\in I_k} \frac{1}{l^\chi} - 1, 1 - \lambda_{min} \sum_{l\in I_k} \frac{1}{l^\chi} \right\} \\ &\leq 1 - \lambda_{min} a' + \epsilon \qquad \text{for all} \quad k \geq K_\epsilon. \end{aligned} \tag{3.40}$$

Moreover, we can use Lemma 3.3 of Appendix, (3.28), (3.27), (3.38), (3.39), Taylor's theorem and the fact $d\|A\|a' < 1$ and to obtain a $K'_\epsilon \geq K_\epsilon$ such that

$$\begin{aligned}\sum_{m=2}^{n_{k+1}-n_k} \frac{\left(\sum_{l\in I_k}\left(\frac{\|\hat{A}_l\|}{l^\chi}+\eta_l\right)\right)^m}{m!} &\leq \sum_{m=2}^{n_{k+1}-n_k} \frac{\left(d\|A\|\sum_{l\in I_k}\frac{1}{l^\chi}+d\|\sum_{l\in I_k}\frac{Y_l}{l^\chi}\|+\sum_{l\in I_k}\eta_l\right)^m}{m!} \\ &\leq \ e^{1+3\epsilon}\frac{(d\|A\|a'+3\epsilon)^2}{2} \qquad \text{for all } k \geq K'_\epsilon. \quad (3.41)\end{aligned}$$

Therefore, by (3.40), Lemma 3.2 iii), (3.36) and (3.41) one finds

$$\begin{aligned}\|U_k\| &\leq \left\|I - A\sum_{l\in I_k}\frac{1}{l^\chi}\right\| + \sum_{l\in I_k}\eta_l + \left\|\sum_{l\in I_k}\frac{Y_l}{l^\chi}\right\| + \sum_{m=2}^{n_{k+1}-n_k}\frac{\left(\sum_{l\in I_k}\left(\frac{\|\hat{A}_l\|}{l^\chi}+\eta_l\right)\right)^m}{m!} \\ &\leq 1 - \lambda_{min}a' + 3\epsilon + e^{1+3\epsilon}\frac{(d\|A\|a'+3\epsilon)^2}{2} \qquad \forall\ k \geq K'_\epsilon. \qquad (3.42)\end{aligned}$$

Furthermore, using the fact that $a' < \frac{\lambda_{min}}{e^1(d\|A\|)^2}$ and making for $\epsilon > 0$ small enough, we find from (3.42) that, there exists a $0 < \gamma < 1$ and an integer $k_1 > 0$ such that

$$\|U_k\| \leq \gamma \qquad \text{for all } k \geq k_1. \qquad (3.43)$$

Step 3: Convergence of remainder w_n along a subsequence using block stability of A_k.

By (3.26), (3.28), (3.30) and $w_k := \nu_k - u_k \to 0$

$$w_{n+1} = \left(I - \frac{\hat{A}_n}{n^\chi} + \eta_n I\right)w_n - \frac{1}{n^\chi}Y_n u_n \quad \text{for } n = 1, 2, \cdots \qquad (3.44)$$

so it follows by (3.44) that

$$\begin{aligned} w_n &= \prod_{l=n_k}^{n-1}\left(I - \frac{\hat{A}_l}{l^\chi} + \eta_l I\right) w_{n_k} \\ &- \sum_{j=n_k}^{n-1}\prod_{l=j+1}^{n-1}\left(I - \frac{\hat{A}_l}{l^\chi} + \eta_l I\right)\frac{Y_j u_j}{j^\chi} \quad \forall \quad n \geq n_k. \end{aligned} \tag{3.45}$$

In particular,

$$w_{n_{k+1}} = U_k w_{n_k} - \sum_{j\in I_k} V_{j,k} u_j \quad \text{for} \quad k = 0, 1, \cdots, \tag{3.46}$$

where U_k is defined in (3.34) and

$$V_{j,k} = \prod_{l=j+1}^{n_{k+1}-1}\left(I - \frac{\hat{A}_l}{l^\chi} + \eta_l I\right)\frac{1}{j^\chi} Y_j. \tag{3.47}$$

By Lemma 3.2 v) and (3.47) we obtain,

$$\begin{aligned} \|V_{j,k}\| &\leq \prod_{l=j+1}^{n_{k+1}-1}\left\|\left(I - \frac{\hat{A}_l}{l^\chi} + \eta_l I\right)\right\| \frac{\|Y_j\|}{j^\chi} \\ &\leq \prod_{l\in I_k}\left(1 + \frac{\|\hat{A}_l\|}{l^\chi} + \eta_l\right)\frac{\|Y_j\|}{j^\chi} \overset{j,k}{\ll} \frac{\|Y_j\|}{j^\chi} \text{ for } j \in I_k,\ k = 0, 1, \ldots \end{aligned} \tag{3.48}$$

Therefore, by (3.43), (3.48), (3.34), (3.46), and (3.28) we have

$$\mid w_{n_k} \mid \overset{k}{\ll} \gamma^{k-k_1} \mid w_{n_{k_1}} \mid + \sum_{l=k_1}^{k-1} \gamma^{k-l-1} \sum_{j\in I_l} \frac{\|A\| + \|\hat{A}_j\|}{j^\chi} \mid u_j \mid \ \forall\, k \geq k_1. \tag{3.49}$$

In addition,

$$\sum_{j\in I_l} \frac{\|A\| + \|\hat{A}_j\|}{j^\chi} \mid u_j \mid = \|A\| \sum_{j\in I_l} \frac{1}{j^\chi} \mid u_j \mid + \sum_{j\in I_l} \frac{\|\hat{A}_j\|}{j^\chi} \mid u_j \mid,$$

so using Lemma 3.2 iv), (3.32), (3.38) and finally applying Toeplitz Lemma, we obtain

$$\lim_{l\to\infty}\sum_{j\in I_l}\frac{\|A\|+\|\hat{A}_j\|}{j^\chi}\mid u_j\mid=0. \tag{3.50}$$

Moreover, since

$$\sum_{l=k_1}^{k-1}\gamma^{k-l-1}=\frac{1-\gamma^{k-k_1}}{1-\gamma}\overset{k}{\ll}1 \qquad \text{for all } k=k_1,k_1+1,\cdots \tag{3.51}$$

it follows from (3.49), (3.50), (3.51) and the Toeplitz Lemma with $a_{l,k}=\gamma^{k-l-1}1_{k_1\le l\le k-1}$ and $x_l=\sum_{j\in I_l}\frac{\|A\|+\|\hat{A}_j\|}{j^\chi}\mid u_j\mid$ that

$$\begin{aligned}\lim_{k\to\infty}\mid w_{n_k}\mid &\le \lim_{k\to\infty}\gamma^{k-k_1}\mid w_{n_{k_1}}\mid\\ &+\lim_{k\to\infty}\sum_{l=k_1}^{k-1}\gamma^{k-l-1}\sum_{j\in I_l}\frac{\|A\|+\|\hat{A}_j\|}{j^\chi}\mid u_j\mid=0.\end{aligned} \tag{3.52}$$

Step 4: Use $w_{n_k}\to 0$ to show block convergence $\max_{n\in I_k}|w_n|\to 0$.

Now, we return to (3.45) and find for $n\in I_k$

$$\begin{aligned}|w_n| &\le \prod_{l=n_k}^{n-1}\left(1+\frac{\|\hat{A}_l\|}{l^\chi}+\eta_l\right)|w_{n_k}|+\sum_{j=n_k}^{n-1}\prod_{l=n_k}^{n-1}\left(1+\frac{\|\hat{A}_l\|}{l^\chi}+\eta_l\right)\frac{\|Y_j\|}{j^\chi}|u_j|\\ &\le \prod_{l\in I_k}\left(1+\frac{\|\hat{A}_l\|}{l^\chi}+\eta_l\right)\left\{|w_{n_k}|+\sum_{j\in I_k}\frac{\|Y_j\|}{j^\chi}|u_j|\right\}\\ &\le \prod_{l\in I_k}\left(1+\frac{\|\hat{A}_l\|}{l^\chi}+\eta_l\right)\left\{|w_{n_k}|+\sum_{j\in I_k}\frac{\|\hat{A}_j\|+\|A\|}{j^\chi}|u_j|\right\}.\end{aligned} \tag{3.53}$$

Finally, by (3.53), (3.52), Lemma 3.2 v), and (3.50) we obtain

$$\lim_{k\to\infty}\max_{n\in I_k}\mid w_n\mid=0. \qquad \Box \tag{3.54}$$

Part b) By (3.13) and (3.28), $z_k = k^\chi(\nu_{k+1} - \nu_k) + \bar{A}_k \nu_k$. Averaging, then reordering the sum, we have

$$\begin{aligned}
\frac{1}{n^\chi}\sum_{k=1}^{n} z_k &= \frac{1}{n^\chi}\left(\sum_{k=1}^{n} k^\chi(\nu_{k+1}-\nu_k) + \sum_{k=1}^{n}\bar{A}_k\nu_k\right) \\
&= \nu_{n+1} - \frac{1}{n^\chi}\sum_{k=1}^{n}(k^\chi - (k-1)^\chi)\nu_k + \frac{1}{n^\chi}\sum_{k=1}^{n}\bar{A}_k\nu_k
\end{aligned}$$

so

$$\begin{aligned}
\left|\frac{1}{n^\chi}\sum_{k=1}^{n} z_k\right| &\le \mid \nu_{n+1} \mid + \sum_{k=1}^{n}\frac{k^\chi - (k-1)^\chi)}{n^\chi} \mid \nu_k \mid \\
&+ \sum_{k=1}^{n}\frac{k^{\chi-1}}{n^\chi}\|\bar{A}_k\| k^{1-\chi} \mid \nu_k \mid . \qquad (3.55)
\end{aligned}$$

The second and third terms on the RHS of (3.55) converge to 0 by the Toeplitz lemma with $a_{n,k} = \frac{k^\chi-(k-1)^\chi}{n^\chi}$, $x_k = |\nu_k|$ and with $a_{n,k} = \frac{k^{\chi-1}\|\bar{A}_k\|}{n^\chi}$, $x_k = k^{\chi-1}|\nu_k|$ respectively. □

3.7 Appendix

We first establish our promised comparison on our conditions.

Lemma 3.1 $\limsup_{n\to\infty}\left\|\frac{1}{n^\chi}\sum_{k=1}^{n}(\bar{A}_k - A)\right\| = 0$ *implies* $\frac{1}{n^\chi}\sum_{k=1}^{n} k^{\chi-1}\|\bar{A}_k\|$ *is bounded in* n.

Proof. By Lemma 3.3 (to follow) and the fact that $\sum_{k=1}^{n} k^{\chi-1} \le \frac{n^\chi}{\chi}$, one finds that

$$\begin{aligned}\frac{1}{n^\chi}\sum_{k=1}^n k^{\chi-1}\|\bar{A}_k\| &\leq \frac{d}{n^\chi}\left\|\sum_{k=1}^n k^{\chi-1}\bar{A}_k\right\| \\ &\leq \frac{d}{n^\chi}\left\|\sum_{k=1}^n k^{\chi-1}(\bar{A}_k - A)\right\| + \frac{d}{n^\chi}\|A\|\sum_{k=1}^n k^{\chi-1} \\ &\leq \frac{d}{n^\chi}\left\|\sum_{k=1}^n k^{\chi-1}(\bar{A}_k - A)\right\| + \frac{d\|A\|}{\chi}. \qquad (3.56)\end{aligned}$$

Noting $\sum_{j=2}^{k}(j^{\chi-1}-(j-1)^{\chi-1}) = k^{\chi-1}-1$, setting $C = \frac{d\|A\|}{\chi} + \sup_n \frac{d}{n^\chi}\left\|\sum_{k=1}^n(\bar{A}_k - A)\right\| < \infty$ and interchanging summation order, we have

$$\begin{aligned}\frac{1}{n^\chi}\sum_{k=1}^n k^{\chi-1}\|\bar{A}_k\| &\leq \frac{d}{n^\chi}\left\|\sum_{k=2}^n\sum_{j=2}^k (j^{\chi-1}-(j-1)^{\chi-1})(\bar{A}_k - A)\right\| + C \\ &\leq \frac{d}{n^\chi}\left\|\sum_{j=2}^n (j^{\chi-1}-(j-1)^{\chi-1})\sum_{k=j}^n(\bar{A}_k - A)\right\| + C. (3.57)\end{aligned}$$

However, by Taylor's theorem $(j^{1-\chi}-(j-1)^{1-\chi}) \leq (1-\chi)(j-1)^{-\chi}$, so by (3.57) we have

$$\begin{aligned}&\frac{1}{n^\chi}\sum_{k=1}^n k^{\chi-1}\|\bar{A}_k\| \\ \leq\ & d\sum_{j=2}^n \frac{(j^{1-\chi}-(j-1)^{1-\chi})}{j^{1-\chi}(j-1)^{1-\chi}}\cdot\frac{1}{n^\chi}\left\|\sum_{k=j}^n(\bar{A}_k - A)\right\| + C \\ \leq\ & d\sum_{j=2}^n \frac{(j-1)^{-\chi}(1-\chi)}{j^{1-\chi}(j-1)^{1-\chi}}.\ \frac{1}{n^\chi}\left(\left\|\sum_{k=1}^n(\bar{A}_k - A)\right\| + \left\|\sum_{k=1}^{j-1}(\bar{A}_k - A)\right\|\right) + C \\ \leq\ & 2d(1-\chi)\sum_{j=2}^n\frac{1}{j^{2-\chi}}\left(\left\|\frac{1}{n^\chi}\sum_{k=1}^n(\bar{A}_k - A)\right\| + \left\|\frac{1}{(j-1)^\chi}\sum_{k=1}^{j-1}(\bar{A}_k - A)\right\|\right) + C\end{aligned}$$

This final term is bounded by the Toeplitz lemma and our hypothesis. □

We give our list of technical bounds used in the proof of Proposition 3.1.

Lemma 3.2 *Assume the setting of Proposition 3.1; and $F_{j,k}$, $\bar{F}_{j,k}$, I_k, $\{z_k\}_{k=1}^{\infty}$ and $\{Y_k\}_{k=1}^{\infty}$ are as defined in (3.33), (3.31) and (3.28). Then, following are true:*

i) $\displaystyle\lim_{n\to\infty}\left\|\prod_{l=1}^{n-1}\left(I-\frac{A}{l^{\chi}}+\eta_l I\right)\right\|=0$

ii) $\displaystyle\lim_{n\to\infty}\left|\sum_{j=1}^{n-1}F_{j,n}z_j\right|=0$ *and* $\displaystyle\lim_{n\to\infty}\left|\sum_{j=1}^{n-1}\bar{F}_{j,n}h\right|=0$

iii) $\displaystyle\lim_{k\to\infty}\left\|\sum_{l\in I_k}\frac{Y_l}{l^{\chi}}\right\|=0$

iv) $\displaystyle\sum_{l\in I_k}\left(\frac{\|\hat{A}_l\|}{l^{\chi}}+\eta_l\right)\overset{k}{\ll}1$ *for all* $k=0,1,\cdots$

v) $\displaystyle\prod_{l\in I_k}\left(1+\frac{\|\hat{A}_l\|}{l^{\chi}}+\eta_l\right)\overset{k}{\ll}1$ *for all* $k=0,1,\cdots$

Proof. i) We know $\left\|I-\frac{A}{l^{\chi}}+\eta_l I\right\|$ is the maximum eigenvalue of $\left((1+\eta_l)I-\frac{A}{l^{\chi}}\right)$ and

$$0\le\left\|\prod_{l=1}^{n-1}\left((1+\eta_l)I-\frac{A}{l^{\chi}}\right)\right\|\le\prod_{l=1}^{n-1}\left\|(1+\eta_l)I-\frac{A}{l^{\chi}}\right\|.$$

Let $\lambda_{min}>0$ be the minimum eigenvalue of A; recall from the statement of Proposition 3.1 that $\eta_k\le\frac{\bar{\eta}}{k^{\theta}}$ and $\theta>\chi$; and fix l^* large enough that: $1+\eta_l-\frac{\lambda_{min}}{l^{\chi}}>0\,\forall\,l>l^*$. Using the fact that $\displaystyle\prod_l(1+x_l)\le\exp\left(\sum_l x_l\right)$, one finds

$$\begin{aligned}
\prod_{l=l^*}^{n-1}\left\|(1+\eta_l)I-\frac{A}{l^\chi}\right\| &\leq \prod_{l=l^*}^{n-1}\left(1+\frac{\bar{\eta}}{l^\theta}-\frac{\lambda_{min}}{l^\chi}\right)\\
&\leq \exp\left(\int_{l^*-1}^{n-1}\frac{\bar{\eta}}{x^\theta}dx-\int_{l^*}^{n}\frac{\lambda_{min}}{x^\chi}dx\right)\\
&\leq \exp\left(D+\frac{\bar{\eta}}{1-\theta}(n-1)^{1-\theta}-\frac{\lambda_{min}}{1-\chi}n^{1-\chi}\right)\\
&\overset{n}{\ll} \exp\left(\frac{-\lambda_{min}}{2-2\chi}n^{1-\chi}\right)
\end{aligned}$$

for some $D\in\mathbb{R}$. Hence,

$$\prod_{l=l^*}^{n-1}\left\|(1+\eta_l)I-\frac{A}{l^\chi}\right\|\to 0 \quad \text{as} \quad n\to\infty. \tag{3.58}$$

ii) $\|(r^\chi+\eta_r r^\chi-(r-1)^\chi)I-A\| \leq\mid (r^\chi-(r-1)^\chi)\mid +\bar{\eta}r^{\chi-\theta}+\|A\|\leq 1+\bar{\eta}+\|A\|$ is upper bounded $\forall r>1$ since $\chi\in(0,1)$. Hence, by (3.31) we have

$$\begin{aligned}
\|F_{r-1,n}-F_{r,n}\| = &\left\|\prod_{l=r+1}^{n-1}\left((1+\eta_l)I-\frac{A}{l^\chi}\right)\left[\frac{1}{(r-1)^\chi}\left((1+\eta_r)I-\frac{A}{r^\chi}\right)-\frac{1}{r^\chi}I\right]\right\|\\
\leq &\left\|\prod_{l=r+1}^{n-1}\left((1+\eta_l)I-\frac{A}{l^\chi}\right)\right\|\frac{1}{r^\chi(r-1)^\chi}\\
\times &\ \|(r^\chi+\eta_r r^\chi-(r-1)^\chi)I-A\|\\
\overset{r,n}{\ll} &\ \frac{1}{r^\chi(r-1)^\chi}\left\|\prod_{l=r+1}^{n-1}\left((1+\eta_l)I-\frac{A}{l^\chi}\right)\right\|
\end{aligned} \tag{3.59}$$

for all $r=2,3,...,n-1$, $n=3,4,....$ Letting λ denote an arbitrary eigenvalue of A, setting $L^c=\left\{l:\frac{\lambda}{l^\chi}-1-\eta_l\geq c\right\}$, noting that $\left\|\left(1+\eta_l-\frac{\lambda}{l^\chi}\right)I\right\|\leq \left(\frac{\lambda}{l^\chi}-1-\eta_l\right)\vee\exp\left(\eta_l-\frac{\lambda}{l^\chi}\right)$ and defining constant $C\doteq\prod_{l\in L^1}\left(\frac{\lambda}{l^\chi}-1-\eta_l\right)\times$

$\exp\left(\sum_{l\in L^0}\frac{\lambda}{l^\chi}-\eta_l\right)$ we have that

$$\begin{aligned}\left\|\prod_{l=r+1}^{n-1}\left(1+\eta_l-\frac{\lambda}{l^\chi}\right)I\right\| &\le \prod_{l\in L^1}\left(\frac{\lambda}{l^\chi}-1-\eta_l\right)\times\exp\left(\sum_{l=r+1,l\notin L^0}^{n-1}\eta_l-\frac{\lambda}{l^\chi}\right)\\ &\le C\exp\left(\sum_{l=r+1}^{n-1}\frac{\bar\eta}{l^\theta}-\frac{\lambda}{l^\chi}\right)\\ &\overset{r,n}{\ll} \exp\left(-\frac{\lambda_{min}}{2-2\chi}\{n^{1-\chi}-(r+1)^{1-\chi}\}\right) \qquad (3.60)\end{aligned}$$

and it follows from (3.60), the fact that the eigenvectors of A span $\mathbb{R}^d$ and the principle of uniform boundedness that

$$\left\|\prod_{l=r+1}^{n-1}\left((1+\eta_l)I-\frac{A}{l^\chi}\right)\right\|\overset{r,n}{\ll} e^{-\frac{\lambda_{min}}{2-2\chi}\{n^{1-\chi}-(r+1)^{1-\chi}\}}. \qquad (3.61)$$

It follows by (3.31), (3.59) and (3.61) that

$$\begin{aligned}\sum_{r=2}^{n-1}(r-1)^\chi\|F_{r-1,n}-F_{r,n}\| &\overset{n}{\ll} \sum_{r=2}^{n-1}\frac{1}{r^\chi}e^{-\frac{\lambda_{min}}{2-2\chi}\{n^{1-\chi}-(r+1)^{1-\chi}\}}\\ &\overset{n}{\ll} e^{-\frac{\lambda_{min}}{2-2\chi}n^{1-\chi}}\int_2^n\frac{1}{t^\chi}e^{\frac{\lambda_{min}}{2-2\chi}t^{1-\chi}}dt\\ &\overset{n}{\ll} 1 \quad \forall\, n=3,4,... \qquad (3.62)\end{aligned}$$

Next, $\sum_{j=1}^{n-1}F_{j,n}z_j=\sum_{j=1}^{n-1}F_{n-1,n}z_j+\sum_{j=1}^{n-1}\left(\sum_{r=j+1}^{n-1}F_{r-1,n}-F_{r,n}\right)z_j$ and

$$\left|\sum_{j=1}^{n-1}\sum_{r=j+1}^{n-1}(F_{r-1,n}-F_{r,n})z_j\right| \le \sum_{r=2}^{n-1}\|F_{r-1,n}-F_{r,n}\|\left|\sum_{j=1}^{r-1}z_j\right|.$$

Therefore, by assumption, (3.31), (3.62) and Toeplitz' lemma with

$$x_r = \frac{1}{(r-1)^\chi} \mid \sum_{j=1}^{r-1} z_j \mid \text{ and } a_{n,r} = (r-1)^\chi \|F_{r-1,n} - F_{r,n}\|$$

we have:

$$\begin{aligned} \left|\sum_{j=1}^{n-1} F_{j,n} z_j\right| &\leq \|F_{n-1,n}\| \left|\sum_{j=1}^{n-1} z_j\right| + \left|\sum_{j=1}^{n-1} \sum_{r=j+1}^{n-1} (F_{r-1,n} - F_{r,n}) z_j\right| \\ &\leq \frac{1}{(n-1)^\chi} \left|\sum_{j=1}^{n-1} z_j\right| + \sum_{r=2}^{n-1} \|F_{r-1,n} - F_{r,n}\| \left|\sum_{j=1}^{r-1} z_j\right| \to 0.(3.63) \end{aligned}$$

as $n \to \infty$.Turning to the second limit in ii), we have by (3.31) and (3.61) that

$$\sum_{j=1}^{n-1} \|\bar{F}_{j,n}\| \overset{n}{\ll} \sum_{j=1}^{n-1} j^{-\chi} e^{-\frac{\lambda_{min}}{2-2\chi}\{n^{1-\chi}-(j+1)^{1-\chi}\}} j^{\chi-\theta}. \tag{3.64}$$

However,

$$\begin{aligned} \sum_{j=1}^{n-1} \frac{1}{j^\chi} e^{-\frac{\lambda_{min}}{2-2\chi}\{n^{1-\chi}-(j+1)^{1-\chi}\}} &\overset{n}{\ll} e^{-\frac{\lambda_{min}}{2-2\chi} n^{1-\chi}} \int_1^n \frac{1}{t^\chi} e^{\frac{\lambda_{min}}{2-2\chi} t^{1-\chi}} dt \\ &\overset{n}{\ll} 1 \end{aligned} \tag{3.65}$$

for all n so the second limit in ii) follows by the Toeplitz lemma.

iii) Since

$$\frac{1}{l^\chi} = \frac{1}{n_k^\chi} + \sum_{r=n_k}^{l-1} \left(\frac{1}{(r+1)^\chi} - \frac{1}{r^\chi}\right) \qquad \forall \quad l \in I_k,$$

one has that

$$\left\|\sum_{l\in I_k} \frac{Y_l}{l^\chi}\right\| \leq \frac{1}{n_k^\chi} \left\|\sum_{l\in I_k} Y_l\right\| + \left\|\sum_{\substack{r<l \\ r,l\in I_k}} \left(\frac{1}{(r+1)^\chi} - \frac{1}{r^\chi}\right) Y_l\right\|.$$

Hence, by Taylor's theorem

$$\begin{aligned}
\left\|\sum_{l\in I_k}\frac{Y_l}{l^\chi}\right\| &\leq \frac{1}{n_k^\chi}\left(\left\|\sum_{l<n_{k+1}} Y_l\right\| + \left\|\sum_{l<n_k} Y_l\right\|\right)\\
&+ \sum_{r=n_k}^{n_{k+1}-2}\frac{r^\chi-(r+1)^\chi}{r^\chi(r+1)^\chi}\left\|\sum_{l<n_{k+1}} Y_l - \sum_{l\leq r} Y_l\right\|\\
&\leq \frac{1}{n_k^\chi}\left(\left\|\sum_{l<n_{k+1}} Y_l\right\| + \left\|\sum_{l<n_k} Y_l\right\|\right)\\
&+ \sum_{r\in I_k}\frac{\chi}{r^{\chi+1}}\left(\left\|\sum_{l<n_{k+1}} Y_l\right\| + \left\|\sum_{l\leq r} Y_l\right\|\right), \qquad (3.66)
\end{aligned}$$

where the summations all start from $l=1$ and stop at $l=n_k-1$, r or $n_{k+1}-1$. Furthermore, by the hypothesis and (3.33) we have that

$$\lim_{k\to\infty}\max_{r\in I_k}\frac{1}{r^\chi}\left\|\sum_{l<n_{k+1}} Y_l\right\| = 0 \qquad (3.67)$$

and the first two terms on the RHS of (3.66) go to zero. Moreover, by (3.33)

$$\begin{aligned}
\sum_{r\in I_k}\frac{1}{r} &\leq \log\left(\frac{n_{k+1}-1}{n_k-1}\right)\\
&= \log\left(\frac{\lfloor(a(k+1))^{\frac{1}{1-\chi}}\rfloor-1}{\lfloor(ak)^{\frac{1}{1-\chi}}\rfloor-1}\right)\to 0\,\text{as}\quad k\to\infty \qquad (3.68)
\end{aligned}$$

due to the fact that

$$1\leq\frac{\lfloor(a(k+1))^{\frac{1}{1-\chi}}\rfloor-1}{\lfloor(ak)^{\frac{1}{1-\chi}}\rfloor-1}\leq\frac{(a(k+1))^{\frac{1}{1-\chi}}}{(ak)^{\frac{1}{1-\chi}}-2}=\frac{(\frac{k+1}{k})^{\frac{1}{1-\chi}}}{1-(\frac{2}{ak})^{\frac{1}{1-\chi}}}\to 1\,\text{as}\quad k\to\infty.$$

In addition, by assumption, (3.67) and (3.68)

$$\sum_{r\in I_k}\frac{\chi}{r^{\chi+1}}\left\|\sum_{l<n_{k+1}}Y_l\right\|\overset{k}{\ll}\sum_{r\in I_k}\frac{1}{r}\frac{1}{n_k^{\chi}}\left\|\sum_{l<n_{k+1}}Y_l\right\|\to 0 \text{ as } k\to\infty$$

and

$$\sum_{r\in I_k}\frac{\chi}{r^{\chi+1}}\left\|\sum_{l\le r}Y_l\right\|\le\sum_{r\in I_k}\frac{\chi}{r}\frac{1}{r^{\chi}}\left\|\sum_{l\le r}Y_l\right\|\to 0 \text{ as } k\to\infty$$

Hence, the last term on the RHS of (3.66) goes to zero too.

iv) By Lemma 3.3, the fact that $\|B\|\le|||\ B\ |||\le\sqrt{d}\|B\|$ for a matrix with rank d, iii) and (3.38) we have

$$\begin{aligned}\sum_{l\in I_k}\left\|\frac{\hat{A}_l}{l^{\chi}}\right\| &\le \sum_{l\in I_k}\left|\left|\left|\frac{\hat{A}_l}{l^{\chi}}\right|\right|\right|\le\sqrt{d}\left|\left|\left|\sum_{l\in I_k}\frac{\hat{A}_l}{l^{\chi}}\right|\right|\right|\le d\left\|\sum_{l\in I_k}\frac{\hat{A}_l}{l^{\chi}}\right\| \\ &\le d\left\|\sum_{l\in I_k}\frac{(\hat{A}_l-A)}{l^{\chi}}\right\|+d\|A\|\sum_{l\in I_k}\frac{1}{l^{\chi}} \\ &\le d\left\|\sum_{l\in I_k}\frac{Y_l}{l^{\chi}}\right\|+d\|A\|\sum_{l\in I_k}\frac{1}{l^{\chi}}\overset{k}{\ll}1 \quad \text{for } k=0,1,2,... \quad (3.69)\end{aligned}$$

Moreover, by (3.39) $\sum_{l\in I_k}\eta_l\le\bar{\eta}\sum_{l\in I_k}\frac{1}{l^{\theta}}\to 0$ as $k\to\infty$.

v) This follows by iv) and the fact that

$$\begin{aligned}\prod_{l\in I_k}\left(1+\frac{\|\hat{A}_l\|}{l^{\chi}}+\eta_l\right) &\le \exp\left(\sum_{l\in I_k}\frac{\|\hat{A}_l\|}{l^{\chi}}+\eta_l\right) \\ &\overset{k}{\ll} 1, \quad \forall\, k=0,1,...\ \square \qquad (3.70)\end{aligned}$$

The following lemma is taken from Kouritzin [21].

Lemma 3.3 *Suppose m is a positive integer and $\{M_k, k = 1, 2, 3, ...\}$ is a sequence of symmetric, positive semidefinite $R^{m\times m}$-matrices. Then, it follows that*

$$\sum_{k=1}^{j} ||| M_k ||| \leq \sqrt{m} \left|\left|\left| \sum_{k=1}^{j} M_k \right|\right|\right|, \qquad \forall j = 1, 2, 3, ...$$

Bibliography

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

[2] Berbee, H. (1987). *Convergence rates in the strong law for a bounded mixing sequence.* Probability Theory and Related Fields, **vol**. 74, pp. 253-270.

[3] Bertsekas, D.P. and Tsitsiklis, J.N. (1996). *Neuro-Dynamic Programming.* Atlanta, GA: Athena Scientific.

[4] Berger, E. (1997). *An almost sure invariance principle for stochastic approximation procedures in linear filtering theory.* Ann. Appl. Probab. **vol.** 7, No. 2, pp. 444-459.

[5] Chen, H.F. (1996). *Recent developments in stochastic approximation.* Proc. IFAC World Congr., pp. 375-380.

[6] Chong, E.K.P., Wang, I.J. and Kulkarni, S.R. (1999). *Noise conditions for prespecified convergence rates of stochastic approximation algorithms.* IEEE Trans. Inform. Theory, **vol.** 45, pp. 810-814.

[7] Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). *Power-Law Distributions in Empirical Data.* Journal SIAM Review, **vol.** 51, Issue 4, pp. 661-703.

[8] Delyon, B. (2000). *Stochastic approximation with decreasing gain: Convergence and asymptotic theory.* Tech. report, Universit de Rennes, Rennes, France.

[9] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Berlin, Germany: Springer-Verlag.

[10] Dippon J̈. and Walk, H. (2006). *The Averaged Robbins Monro Method for Linear Problems in a Banach Space.* Journal of Theoretical Probability, **vol.** 19, No. 1, pp. 166-189.

[11] Eweda E. and Macchi, O. (1984). *Convergence of an adaptive linear estimation algorithm.* IEEE Trans. Automat. Contr., **vol.** AC-29, pp. 119-127.

[12] Even-Dar, E. and Mansour, Y. (2004). *Learning rates for q-learning.* Journal of Machine Learning Research, **vol.** 5, pp. 1-25.

[13] Farden, D.C. (1981). *Stochastic Approximation with Correlated Data.* IEEE Trans. Inform. Theory, **vol.** IT-27, NO. 1, pp. 105-113.

[14] Frost, O.L. (1972). *An algorithm for linearly constrained adaptive array processing.* Proc. IEEE, **vol.** 60, pp. 922-935.

[15] George, A.P. and Powell, W. B. (2006). *Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming.* Journal of Machine Learning Research, **vol.** 65, pp. 167-198.

[16] Griffiths, L.J. (1969). *A simple algorithm for real-time processing in antenna arrays.* Proc. IEEE, **vol.** 57, pp. 1696-1704.

[17] Györfi, L. (1980). *Stochastic approximation from ergodic sample for linear regression.* Z. Wahrscheinlichkeitstheorie und verwandte Gebiete, **vol.** 54, pp. 47-55.

[18] Györfi, L. (1984). *Adaptive linear procedures under general conditions.* IEEE Trans. Inform. Theory, **vol.** IT-30, pp. 262-267.

[19] Karagiannis, T. , Molle, M. and Faloutsos, M. (2004). *Long-Range Dependence Ten Years of Internet Traffic Modeling.* IEEE Computer Society, **vol.** 8, No. 5, pp. 57-64.

[20] Kouritzin, M.A. and Sadeghi, S. (2015). *Marcinkiewicz Law of Large Numbers for Outer-products of Heavy-tailed, Long-range-Dependence Data.* Advances in Applied Probability Journal, in press.

[21] Kouritzin, M.A. (1996). *On the convergence of linear stochastic approximation procedures.* IEEE Trans. Inform. Theory, **vol.** 42, pp. 1305-1309.

[22] Kouritzin, M.A. (1996). *On the interrelation of almost sure invariance principles for certain stochastic adaptive algorithms and for partial sums of random variables.* Journal of Theoretical Probability, **vol.** 9, No. 4, pp. 811-840.

[23] Kouritzin, M.A. (1994). *On Almost-Sure Bounds for the LMS Algorithm.* IEEE Trans. Inform. Theory, **vol.** 40, No. 2, pp. 372-383.

[24] Kouritzin, M.A. (1994). *Inductive methods and rates of r-mean convergence in adaptive filtering.* Stochastics and Stochastics Reports, **vol.** 51, Issue 3-4, pp. 241-266.

[25] Kushner, H. J. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications.* Springer, Second edition, pp. 8.

[26] Ljung, L., Pflug, G. and Walk, H. (1992). *Stochastic Approximation and Optimization of Random Systems.* Basel, Switzerland: Birkhäuser-Verlag.

[27] Ljung, L. (1999). *System Identification: Theory for the User.* Prentice Hall, second edition.

[28] Louhchi, S. and Soulier, P. (2000). *Marcinkiewicz-Zegmond Strong Laws for Infinite Variance Time Series.* Statistical Inference for Stochastic Processes, **vol**. 3, pp. 31-40.

[29] Rio, E. (1995). *A Maximal Inequality and Dependent Marcinkiewicz-Zegmond Strong Laws.* The Annals of Probability, **vol.** 23, No. 2, pp. 918-937.

[30] Robbins, H. and Monro, S. (1951). *A stochastic approximation method.* Ann. Math. statist., **vol.** 22, pp. 400-407.

[31] Soderstrom, T. and Stoica, P. (1989). *System Identification.* Prentice Hall.

[32] Solo, V. and Kong, X. (1995). *Adaptive Signal Processing Algorithms: Stability and Performance.* Englewood Cliffs, NJ: Prentice-Hall.

[33] Stout, W.F. (1974). *Almost Sure Convergence.* Academic Press Inc., pp. 126.

[34] Tadić, V.B. (2004). *On the Almost Sure Rate of Convergence of Linear Stochastic Approximation Algorithms.* IEEE Trans. Inform. Theory, **vol**. 50, No. 2, pp. 401-409.

[35] Thanh, L.V., Yin, G. and Wang, L.Y. (2011). *State observers with random sampling times and convergence analysis of double-indexed and randomly-weighted sums of mixing processes.* SIAM J. Control Optim., **vol.** 49, No. 1, pp. 106-124.

[36] Walk, H. and Zsidó, L. (1989). *Convergence of Robbins-Monro method for linear problems in banach space.* J. Math. Anal. Applic., **vol**. 139, pp. 152-177.

[37] Yin, G. (1992). *Asymptotic Optimal Rate of Convergence for an Adaptive Estimation Procedure.* Stochastic Theory and Adaptive Control, Lecture Notes in Control and Information Sciences, **vol.** 184, pp. 480-489.

Chapter 4

Research Status and General comments

This chapter contains a summary of the research contribution of this thesis and a brief introduction to the authors' future goals. The research proposed in this chapter is related to the results in chapters 2 and 3.

4.1 Current Contribution

In chapter 2 we studied the Marcinkiewicz strong law, for outer products two two-sided linear processes with matrix sequences of coefficient that could decay slowly enough that linear processes demonstrate long-range dependence while the outer product could have heavy tails. In particular, the heavy tail and long-range dependence phenomena for outer products of linear processes were handled simultaneously and a new decoupling property was proved that showed the convergence rate was determined by the worst of the heavy tails or the long-range dependence, but not the combination. The main result was applied to obtain Marcinkiewicz strong law of large numbers for nonlinear functions of partial sums of stochastic processes, autocovariances and stochastic approximation.

In the next chapter almost sure convergence rates for linear stochastic approximation algorithms were studied under the assumptions that were implied by the Marcinkiewicz strong law of large numbers, which allows the coefficients of the stochastic approximation algorithm to have heavy tails, long-range dependence or both. It seems that we are the first to consider processes that are simultaneously heavy-tailed and long-range dependent in stochastic approximation. Finally, corroborating experimental outcomes and decreasing-gain design considerations were provided to verify results experimentally in the stochastic approximation setting.

4.2 Future Research

The next phase of this research will be concentrated on showing the optimality of obtained Marcinkiewicz strong law of large numbers results in previous chapters in a polynomial sense. This means that we cannot achieve a polynomially better rate than what we already obtained by Marcinkiewicz strong law of large numbers results. To achieve this goal, following steps are of interest:

1) Obtain functional non-central limit theorem for outer product of linear processes with heavy-tailed and long-range dependence.

2) Transfer functional non-central limit theorem from the coefficients of stochastic approximation algorithm to its solution.

Hence, the result of immediate interest to the author is at first deriving functional non-central limit theorem for outer product and sample auto-covariance functions of linear processes with long-range dependence and innovation of infinite fourth moment and finally use these results to get functional non-central limit theorem under stochastic approximation setting.

Before describing the mentioned steps in more details, we explain weak types of convergence briefly and give a short literature review.

For statistical inference in time series, it is usually necessary to rely on asymptotic convergence results such as law of large numbers and the central limit theorem. In fact, the central limit theorem is one of the most useful tools for studying the asymptotic distribution of estimators (e.g., kernel type density estimators, econometric estimators). The central limit theorem provides conditions ensuring that the standardized sum of a sequence of random variables has approximately the standard normal distribution, in large samples.

Theorem 4.1 *Suppose $\{Y_1, Y_2, \cdots\}$ is a sequence of i.i.d. random variables with $E[Y_i] = \mu$ and $Var[Y_i] = \sigma^2 < \infty$. Then as n approaches infinity,*

$$\sqrt{n}\Bigg(\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\Big) - \mu\Bigg) \ \overset{d}{\rightarrow}\ Z \sim N(0,\ \sigma^2).$$

A generalization of the central limit theorem is called functional central limit theorem, or weak invariance principle, which has found application in studying the asymptotic distribution of estimators. Doob in 1949 asked whether the convergence in distribution held for more general functionals, thus formulating a problem of weak convergence of random functions in a suitable function space. The functional central limit theorem can be viewed as a generalization of the central limit theorem to metric spaces other than finite-dimensional

Euclidean spaces.

Theorem 4.2 *Suppose* $S_n = \sum_{i=1}^{n} Y_i$ *where* $\{Y_1, Y_2, \cdots\}$ *is a sequence of i.i.d. random variables with* $E[Y_i] = \mu$, $Var[Y_i] = \sigma^2 < \infty$ *and define the stochastic process* $X^{(n)}$ *by*

$$X_t^{(n)} = \frac{S_{\lfloor nt \rfloor} - E[S_{\lfloor nt \rfloor}]}{\sqrt{n}} \qquad \forall\, 0 \leq t \leq 1$$

Then, $X_t^{(n)} \overset{d}{\to} B_t$ *where* B *is Brownian motion with diffusion coefficient* σ.

Finite variance has been an essential property of the deriving process in all the foregoing results which without this, there is no central limit theorem in the usual sense. To get the idea what might happen instead, we require to define the class of stable distributions. A random variable Y has a stable distribution if for any n, there exist independent copies of Y as $Y_1, Y_2, \cdots Y_n$ and real sequences $\{a_n, b_n\}$, where $a_n > 0$, such that $S_n = Y_1 + Y_2 + \cdots + Y_n$ has the same distribution as $a_n Y + b_n$. The distribution is called strictly stable when $b_n = 0$. It turns out that we necessarily have, $a_n = n^{\frac{1}{\alpha}}$ for some $0 < \alpha \leq 2$. Above all, members of the stable class can be negatively or positively skewed around a point of central tendency. Distributions having this property for a_n (not necessarily with $b_n = 0$) are called stable with exponent α or α-stable.

In the recent years, the asymptotic behavior of random variables with heavy tails which arise naturally in various models, for example in financial data, has been investigated increasingly. The oldest result in this direction is the stable limit theorem whose proof can be found in Feller [41]. This theorem states that a suitably normalized sum $S_n = \sum_{i=1}^{n} Y_i$ of a sequence $(Y_i)_{i \geq 1}$ of i.i.d.

random variables with regularly varying tails of index $\alpha \in (0, 2)$ converges in distribution to an α-stable random variable.

Theorem 4.3 *If $S_n = \sum_{i=1}^{n} Y_i$, and $Y_1, Y_2, \cdots Y_n$ are i.i.d. random variables in the domain of attraction of a non-degenerate strictly stable law with parameter α and $E(Y_1) = 0$ for $\alpha > 1$, then*

$$\frac{S_n}{n^{\frac{1}{\alpha}} L(n)} \xrightarrow{d} Z, \quad as \quad n \to \infty$$

for a slowly varying L. Z has a so-called stable distribution.

If convergence to an α-stable law takes the place of the usual central limit theorem, the next step is the existence of a corresponding functional central limit theorem. The functional version of this result is due to Skorokhod [110] and gives the convergence in distribution of the process $\{S_{\lfloor nt \rfloor}\}_{t \geq 0}$ in the space $D[0, 1]$ of cádlág functions on $[0, 1]$ (equipped with the Skorokhod topology J_1) to an α-stable Lévy process.

Note: cádlág is a french acronym standing for "continue á droit, limites á gauche", by way of explanation i.e., functions that may contain jumps, but not isolated points, such as to be discontinuous in both directions. cádlág functions are right continuous with left limits.

Theorem 4.4 (functional non-central limit theorem) *Let X and $\{X^{(n)}, n = 1, 2, , \cdots\}$ be elements of $D_{[0,1]}$. Consider the normalized partial sum process $X^{(n)} \in D_{[0,1]}$ such that*

$$X_t^{(n)} = \frac{1}{n^{\frac{1}{\alpha}} L(n)} \sum_{i=1}^{\lfloor nt \rfloor} Y_i.$$

Then, (under a regularity condition) $X_t^{(n)} \xrightarrow{d} X_t$ as $n \to \infty$, where $\{X_t\}$ is the stable Lévy process with X_1 belonging to α-stable law, and the convergence *holds on the Skorokhod space $D[0,1]$, equipped with the Skorokhod J_1 topology.*

Resnick [100] in 1986 gave a new proof of this result using a powerful idea based on the convergence of a sequence of point processes. Davis and Resnick [24] applied this idea to the more complex situation of linear sequences of the form $X_i = \sum_{j\in\mathbb{Z}} c_j Y_{i-j}$, where $(Y_i)_i$ are i.i.d. random variables with heavy tails. However, Avram and Taqqu [5] showed that this result cannot be extended to a functional convergence result in the Skorokhod topology J_1; on the other hand, if the coefficients $(c_j)_j$ have the same sign, then the functional convergence holds in the Skorokhod topology M_1. Recently, Balan, Jakubowski and Louhichi [6] presented a similar functional convergence result for a linear sequence whose coefficients do not necessarily have the same sign on Skorokhod space equipped with the S topology introduced by Jakubowski [59].

4.2.1 Functional Non-Central Limit Theorem for Outer Product of Linear Processes

The following almost sure convergence under some proper conditions was established in Chapter 2

$$\lim_{n\to\infty} \frac{1}{n^{\frac{1}{p}}} \sum_{k=1}^{n} (D_k - D) = 0 \quad \text{a.s.} \quad \text{for} \quad p < \frac{1}{2-\sigma-\overline{\sigma}} \wedge \alpha \wedge 2,$$

for some $\alpha > 1$ and $(\sigma, \overline{\sigma}) \in \left(\frac{1}{2}, 1\right]$ where $D_k = X_k \overline{X}_k^T$ and $\{X_k\}, \{\overline{X}_k\}$ are both two-sided linear processes. Now, we would like to investigate the weak invariance principle for $\{D_k\}$ when they have heavy tails and long-range dependence. The limit distributions of linear processes generated by i.i.d.

noises have been extensively studied. See, for instance, [3], [24], [25], [62], [5], [55], [54], [129], [7] and [93]. However, the functional convergence of outer products of linear processes with heavy tails and long-range dependence has not been fully investigated in a general form. Hence, we would like to show under proper conditions on coefficients and innovations of linear processes we have following conjecture

a) If $\sigma > 1 - \frac{1}{2\alpha}$, $1 < \alpha < 2$

$$\frac{1}{n^{\frac{1}{\alpha}}} \sum_{k=1}^{\lfloor n. \rfloor} (D_k - D) \xrightarrow{D} Y_.$$

where $\{Y_t\}_{t \in [0,1]}$ are stable processes with index α $(1 < \alpha < 2)$ with sample paths in the space $D_{[0,1]}$ equipped with some appropriate topology (e.g., S topology).

b) If $\frac{1}{2} < \sigma < 1 - \frac{1}{2\alpha}$, $1 < \alpha < 2$

$$\frac{1}{n^{2-2\sigma}} \sum_{k=1}^{\lfloor n. \rfloor} (D_k - D) \xrightarrow{D} C \left[U_\sigma(.) \right]$$

where U_σ is a Rosenblatt process and the convergence is in $D_{[0,1]}$ equipped with J_1 topology.

It is notable that the case a) represents heavy-tailed dominant situation and case b) represents long-range dependence dominant one.

This work is motivated by the paper “Sample autocovariances of long-memory time series ” by Horváth and Kokoszka [54] and “On functional limits of short- and long-memory linear processes with GARCH(1,1) noises” by Zhanga, Sinb and Lingc [133].

Horváth and Kokoszka [54] studied the asymptotic distribution of normalized sample autocovariances of processes with long-memory and infinite fourth moment and proved that the sample autocovariances based on the intensity of dependence converges in distribution to either stable distribution with index α or Rosenblatt process.

On the other hand, Zhanga, Sinb and Lingc [133] investigated the long-memory linear processes but with GARCH (1,1) noises of tail index 2α $(1 < \alpha < 2)$ in a functional sense. They showed that the autocovariances converge to functionals of either α-stable processes or Rosenblatt process again based on how much long-range dependent is present in the model. The weak convergence was established on the space of cádlág functions on [0, 1] with S or J_1 topology.

4.2.2 Stochastic Approximation Algorithm and Functional Non-Central Limit Theorem

Marcinkiewicz strong law of large numbers for stochastic approximation algorithm with decreasing gain has been discussed in Chapter 3. Technically, what was shown was the transfer of Marcinkiewicz strong law of large numbers from the partial sum of linear algorithm's coefficients ($\{A_k\}$ and $\{b_k\}$) with heavy tails or/and long-range dependence to the solution of stochastic approximation algorithms with decreasing gain, $h_{k+1} = h_k + \frac{1}{k^\chi}(b_k - A_k h_k)$. Namely, $n^\gamma|h_n - A^{-1}b| \to 0$ a.s. for the $\gamma < (\chi - \frac{1}{\alpha}) \wedge (\chi + 2\sigma - 2)$ when $\chi \in (0,1)$, $\alpha \in (1,2)$ and $\sigma \in \left(\frac{1}{2}, 1\right]$. Now, we would like to obtain functional non-central limit theorem for the solution of stochastic approximation algorithm in order to demonstrate that attained rate of convergence is optimal in the polynomial sense; which is obtaining weak convergence to a non trivial process at the boundary.

The ultimate goal is deriving functional non-central limit theorem for outer product of linear processes with long-range dependence and heavy tails under stochastic approximation setting. However, the fist step is investigating the transfer method under the sole heavy-tailed phenomenon. This work is motivated by the paper "On the interrelation of almost sure invariance principles for certain stochastic adaptive algorithms and for partial sums of random variables" by Kouritzin [22] and "Functional Convergence of Linear Processes with Heavy-Tailed Innovations" by Balan, Jakubowski and Louhichi [6].

Kouritzin [22] established almost sure invariance principles and showed that if $\{A_k, k = 1, 2, 3, \cdots\}$ and $\{b_k, k = 1, 2, 3, \cdots\}$ are processes satisfying almost sure bounds, then $\{h_k, k = 1, 2, 3, \cdots\}$, the solution of the stochastic approximation adaptive filtering algorithm $h_{k+1} = h_k + \frac{1}{k}(b_k - A_k h_k)$ for $k = 1, 2, 3, \cdots$ also satisfies an almost sure invariance principle of the same type.

On the other hand, Balan, Jakubowski and Louhichi [6] considered the functional convergence of partial sums of linear processes with heavy-tailed innovations and gave necessary and sufficient conditions for the finite dimensional convergence to an α-stable Lévy motion in the case of summable coefficients. However, we have to take into consideration that this work considers the partial sum of linear processes not the outer products of linear processes.

Now, we would like to prove if the normalized sum of coefficients satisfies the functional non-central limit theorem,

$$A^{(n)} \xrightarrow{D} X^A, \; b^{(n)} \xrightarrow{D} X^b$$

on the Skorokhod space $D_{[0,1]}$ equipped with some suitable topology, where

$$A_t^{(n)} = \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{\lfloor nt \rfloor} (A_k - A) \text{ and } b_t^{(n)} = \frac{1}{n^{\chi-\gamma}} \sum_{k=1}^{\lfloor nt \rfloor} (b_k - b), \quad t \in [0,1]$$

with $\{X_t^A, X_t^b\}$ being stable Lévy motion processes with index α $(1 < \alpha < 2)$ and when $\{A_k\}_{k=1}^{\infty}$ are symmetric, positive semidefinite random matrices, A is a positive definite matrix, $\{b_k\}_{k=1}^{\infty}$ are random vector, b is a vector, $\chi \in (0,1)$ and $\gamma = \chi - \frac{1}{\alpha}$ then the solution of above stochastic approximation algorithm with step size $\mu_k = \frac{1}{k^{\chi}}$, also converges on the Skorokhod space $D_{[0,1]}$ to a stable Lévy process with some type of drift. In other word, $H^{(n)} \xrightarrow{D} Y$, where $H_t^{(n)} = n^{\gamma} |h_{\lfloor nt+1 \rfloor} - A^{-1} b|$ and Y_t is a non-symmetric stable Lévy process.

Bibliography

[1] ANDERSON, T.W. AND WALKER, A.M. (1964). *.On the asymptotic distribution of the autocorrelations of a sample from a linear process.* Ann. Math. Statist., **vol.** 35, pp. 1296-1303.

[2] ANDREWS, D.W.K. (1984). *Non-strong mixing autoregressive processes.* Journal of Applied Probability, **vol.** 21, pp. 930-934.

[3] ASTRAUSKAS, A. (1983). *Limit theorems for sums of linearly generated random variables.* Lithuanian Mathematical Journal, **vol.** 23, pp.127-134.

[4] AVRAM, F. AND TAQQU, M.S. (1987). *Generalized powers of strongly dependent random variables.* Ann. Probab., **vol.** 15, pp. 767-775.

[5] AVRAM, F. AND TAQQU, M.S. (1992). *Weak convergence of sums of moving averages in the α-stable domain of attraction.* Lithuanian Math. J., **vol.** 23, pp. 127-134.

[6] BALAN, R.M., JAKUBOWSKI, A. AND LOUHICHI, S. (2014). *Functional Convergence of Linear Processes with Heavy-Tailed Innovations.* Preprint.

[7] BALAN, R.M., JAKUBOWSKI, A. AND LOUHICHI, S. (2012). *Functional convergence of linear sequences in a non-Skorokhod topology.* Preprint.

[8] BENVENISTE, A., MÉTIVIER, M., PRIOURET, P. (1990). *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag.

[9] BERBEE, H. (1987). *Convergence rates in the strong law for a bounded mixing sequence.* Probability Theory and Related Fields, **vol**. 74, pp. 253-270.

[10] Berger, E. (1997). *An almost sure invariance principle for stochastic approximation procedures in linear filtering theory.* Ann. Appl. Probab. **vol.** 7, No. 2, pp. 444-459.

[11] Berkes, I. and Philipp, W. (1979). *Approximation theorems for independent and weakly dependent random vectors.* Ann. Probab., **vol.** 7, pp. 29-54.

[12] Bernoulli, J. (1713). *Ars Conjectandi.* Opus Posthumum, Accedit Tractatus de Seriebus infinitis, et Epistola Gallice scripta de ludo Pilae recticularis, Basileae, (Ch. 1-4 translated into English by Sung B. (1966). *Ars Conjectandi*, Technical Report No. 2, Dept. of Statistics, Harvard University.

[13] Bertsekas, D.P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming.* Atlanta, GA: Athena Scientific.

[14] Borel, E. (1909). *Les probabilites denombrables et leurs applications aritbmttiques.* Rendiconti del Circolo Matematico di Palermo, **vol.** 27, pp. 247-271.

[15] Bradley, R.C. (1983). *Approximation theorems for strongly mixing random variables.* Michigan Math. J., **vol.** 30, pp. 69-81.

[16] Chanda, K.C. (1974). *Strong mixing properties of linear stochastic processes.* Journal of Applied Probability, **vol.** 11, pp. 401-408.

[17] Chandra, T.K. and Ghosal, S. (1996). *Extensions of the Strong Law of Large Numbers of Marcinkiewicz and Zygmund for Dependent Variables.* Acta Math. Hungar., **vol.** 71(4), pp. 327-336.

[18] Chebyshev, P.L. (1867). *Des valuers moyennes.* Journal de Mathématiques Pures et Appliquées, **vol.** 12, pp. 177-184.

[19] Chen, H. (1998b). *Convergence of SA algorithms in multi-root or multi-extreme cases.* Stoch. Stoch. Rep., **vol.** 64, pp. 255-266.

[20] Chernick, M.R. (1981). *A limit theorem for the maximum of autoregressive processes with uniform marginal distributions.* Ann. Prob., **vol.** 9, pp. 145-149.

[21] Chong, E.K.P., Wang, I.J. and Kulkarni, S.R. (1999). *Noise conditions for prespecified convergence rates of stochastic approximation algorithms.* IEEE Trans. Inform. Theory, **vol.** 45, pp. 810-814.

[22] Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). *Power-Law Distributions in Empirical Data.* Journal SIAM Review, **vol.** 51, Issue 4, pp. 661-703.

[23] Dabrowski, A.R., Jakubowski, A. (1993). *Stable limits for associated random variables.* Ann. Probab., **vol.** 1(22), pp. 1-16.

[24] Davis, R.A. and Resnick, S.I. (1985). *Limit theory for moving averages of random variables with regularly varying tail probabilities.* The Annals of Probability, **vol.** 13, pp. 179-195.

[25] Davis, R.A. and Resnick, S.I. (1986). *Limit theory for the sample covariance and correlation functions of moving averages.* Ann. Statist., **vol.** 14, pp. 533-558.

[26] Davis, R.A. and Marengo, J.E. (1990). *Limit theory for sample covariance and correlation matrix functions of a class of multivariate linear processes.* , **vol.** 6, pp. 483-497.

[27] Dedecker, J. and Prieur, C. (2004). *Coupling for τ-dependent sequences and applications.* J. Theoret. Probab., **vol.** 17, pp. 861-885.

[28] Delyon, B. (2000). *Stochastic approximation with decreasing gain: Convergence and asymptotic theory.* Tech. report, Universit de Rennes, Rennes, France.

[29] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Berlin, Germany: Springer-Verlag.

[30] Dippon J̈. and Walk, H. (2006). *The Averaged Robbins Monro Method for Linear Problems in a Banach Space.* Journal of Theoretical Probability, **vol.** 19, No. 1, pp. 166-189.

[31] Dobrushin, R.L. and Major, P. (1979). *Non-central limit theorems for non-linear functions of Gaussian fields.* Z. Wahrscheinlichkeitstheorie Verw. Geb., **vol.** 50, pp. 27-52.

[32] Duflo, M. (1996). *Algorithmes Stochastiques.* Springer.

[33] Doukhan, P. (2003). *Models, inequalities, and limit theorems for stationary sequences.* In Theory and Applications of Long-Range Dependence (P. Doukhan, G. Oppenheim and M.S. Taqqu, eds.), Birkhäuser, Boston, pp. 43-100.

[34] Eberlein, E. and Taqqu, M.S., eds. (1986). *Dependence in Probability and Statistics: A Survey of Recent Results.* Birkhäuser, Boston.

[35] Eberlein, E. (1986). *On strong invariance principles under dependence assumptions.* Ann. Probab., **vol.** 14, pp. 260-270.

[36] Even-Dar, E. and Mansour, Y. (2004). *Learning rates for q-learning.* Journal of Machine Learning Research, **vol.** 5, pp. 1-25.

[37] Eweda E. and Macchi, O. (1984). *Convergence of an adaptive linear estimation algorithm.* IEEE Trans. Automat. Contr., **vol.** AC-29, pp. 119-127.

[38] Fama, E. (1963). *Mandelbrot and the stable Paretian hypothesis.* Journal of Business, **vol.** 36, pp. 420-429.

[39] Fama, E. (1965). *The behavior of stock market prices.* Journal of Business, **vol.** 38, pp. 34-105.

[40] Farden, D.C. (1981). *Stochastic Approximation with Correlated Data.* IEEE Trans. Inform. Theory, **vol.** IT-27, NO. 1, pp. 105-113.

[41] Feller, W. (1911). *An introduction to probability theory and its applications* New York: Wiley **vol.** 2, 3rd edition.

[42] Frost, O.L. (1972). *An algorithm for linearly constrained adaptive array processing.* Proc. IEEE, **vol.** 60, pp. 922-935.

[43] George, A.P. and Powell, W.B. (2006). *Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming.* Journal of Machine Learning Research, **vol.** 65, pp. 167-198.

[44] Giraitis L. and Surgailis, D. (1989). *Limit theorem for polynomials of linear process with long-range dependence.* Lith. Math. J., **vol.** 29, pp. 128-145.

[45] Griffiths, L.J. (1969). *A simple algorithm for real-time processing in antenna arrays.* Proc. IEEE, **vol.** 57, pp. 1696-1704.

[46] Giraitis L. and Surgailis, D. (1986). *Multivariate Appell polynomials and the central limit theorem.* In E. Eberlein and M. S. Taqqu, (eds.), Dependence in Probability and Statistics, Birkhäuser, Boston, pp. 21-71.

[47] Giraitis L. and Surgailis, D. (1990). *A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittles estimate.* Probab. Theory Related Fields, **vol.** 86, pp. 87-104.

[48] Györfi, L. (1980). *Stochastic approximation from ergodic sample for linear regression.* Z. Wahrscheinlichkeitstheorie und verwandte Gebiete, **vol.** 54, pp. 47-55.

[49] Györfi, L. (1984). *Adaptive linear procedures under general conditions.* IEEE Trans. Inform. Theory, **vol.** IT-30, pp. 262-267.

[50] Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

[51] Hall, P. (1997). *On defining and measuring long-range dependence.* Fields Institute Communications, **vol.** 11, pp. 153-160.

[52] Hannan, E.J. (1976). *The asymptotic distribution of serial covariances.* , **vol.** 4, pp. 396-399.

[53] Heyde, C.C., Yang, Y. (1997). *On Defining Long-Range Dependence.* Journal of Applied Probability, **vol.** 34, No. 4, pp. 939-944.

[54] Horváth, L. and Kokoszka, P. (2008). *Sample autocovariances of long-memory time series.* Bernoulli, **vol.** 14, pp. 405-418.

[55] Hosking, J.R.M. (1996). *Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long memory time series.* J. Econometrics, **vol.** 73, pp. 261-284.

[56] Hurst, H. (1951). *Long-term storage capacity of reservoirs.* Transactions of the American Society of Civil Engineers, **vol.** 116, pp. 770-808.

[57] Hurst, H. (1955). *Methods of using long-term storage in reservoirs.* Proceedings of the Institution of Civil Engineers, Part I, pp. 519-577.

[58] Ibragimov, I. A. and Linnik, Y.V. (1971). *Independent and stationary sequences of random variables.* Wolters-Nordhoff,Groningen.

[59] Jakubowski, A. (1997). *A non-Skorohod topology on the Skorohod space.* Electr. J. Probab., **vol.** 2, pp. 1-21.

[60] Karagiannis, T., Molle, M. and Faloutsos, M. (2004). *Long-Range Dependence Ten Years of Internet Traffic Modeling.* IEEE Computer Society, **vol.** 8, No. 5, pp. 57-64.

[61] Karmeshu, D. and Krishnamachari, A. (2004). *Sequence variability and long-range dependence in DNA: An information theoretic perspective.* in Neural Information Processing, pp. 13541361, Berlin: Springer. 3316 of Lecture Notes in Computer Science.

[62] Kasahara, Y., Maejima, M. (1988). *Weighted sums of i.i.d. random variables attracted to integrals of stable processes.* Probab. Theory Related Fields, **vol.** 78, pp. 75-96.

[63] Kolmogorov, A. (1940). *Wienersche Spiralen und einige andere interessante kurven in Hilbertschen raum.* Computes Rendus (Doklady) Academic Sciences USSR (N.S.), **vol.** 26, pp. 115-118.

[64] Kouritzin, M.A. and Sadeghi, S. (2015). *Convergence Rates and Decoupling in Linear Stochastic Approximation Algorithms.* SIAM Journal on Control and Optimization, **vol.** 53-3, pp. 1484-1508.

[65] Kouritzin, M.A. and Sadeghi, S. (2015). *Marcinkiewicz Law of Large Numbers for Outer-products of Heavy-tailed, Long-range-Dependence Data.* Advances in Applied Probability Journal, in press.

[66] Kouritzin, M.A. (1996). *On the convergence of linear stochastic approximation procedures.* IEEE Trans. Inform. Theory, **vol.** 42, pp. 1305-1309.

[67] Kouritzin, M.A. (1996). *On the interrelation of almost sure invariance principles for certain stochastic adaptive algorithms and for partial sums of random variables.* Journal of Theoretical Probability, **vol.** 9, No. 4, pp. 811-840.

[68] Kouritzin, M.A. (1994). *On Almost-Sure Bounds for the LMS Algorithm.* IEEE Trans. Inform. Theory, **vol.** 40, No. 2, pp. 372-383.

[69] Kouritzin, M.A. (1994). *Inductive methods and rates of r-mean convergence in adaptive filtering.* Stochastics and Stochastics Reports, **vol.** 51, Issue 3-4, pp. 241-266.

[70] Kouritzin, M.A. (1995). *Strong approximation for cross-covariances of linear variables with long-range dependence.* Stochastic Processes Appl. **vol.** 60, pp. 343-353.

[71] Kushner, H.J. and Yin, G. (1997). *Stochastic approximation algorithms and applications.* New York: Springer Verlag.

[72] Kushner, H.J. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications.* Springer, Second edition, pp. 8.

[73] Lin, Z. and Lu, C. (1996). *Limit Theory for Mixing Dependent Random Variables.* Kluwer, Dordrecht.

[74] Ljung, L., Soderstrom, T. (1983). *Theory and Practice of Recursive Identification.* MIT Press.

[75] Ljung, L. (1999). *System Identification: Theory for the User.* Prentice Hall, second edition.

[76] Louhichi S. (2000). *Convergence rates in the strong law for associated random variables.* Probab. Math. Stat., **vol.** 20 pp. 203-214.

[77] Louhichi, S. and Rio, E. (2011). *Functional convergence to stable Lévy motions for iterated random Lipschitz mappings.* Electr. J. Probab., **vol.** 16, pp. 2452-2480.

[78] Louhchi, S. and Soulier, P. (2000). *Marcinkiewicz-Zegmond Strong Laws for Infinite Variance Time Series.* Statistical Inference for Stochastic Processes, **vol.** 3, pp. 31-40.

[79] Mandelbrot, B. and Wallis, J. (1968). *Noah, Joseph and operational hydrology.* Water Resources Research, **vol.** 4, pp. 909-918.

[80] Mandelbrot, B. and Van Ness, J. (1968). *Fractional Brownian motions, fractional noises and applications.* SIAM Review, **vol.** 10, pp. 422-437.

[81] Mandelbrot, B. (1963). *New methods in statistical economics.* Journal of Political Economy, **vol.** 71, pp. 421-440.

[82] Mandelbrot, B. (1972). *Statistical methodology for non-periodic cycles: from the covariance to R/S analysis.* Ann. Econ. and Social Measurement, **vol.** 1, pp. 259-290.

[83] Markov, A.A. (1906). *Rasprostranenie zakona bol'shih chisel na velichiny.* zavisyaschie drug ot druga, Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete, 2-ya seriya, **vol.** 15 (94), pp. 135-156.

[84] McCulloch, J.H. (1986). *Simple consistent estimators of stable distribution parameters.* Communications in Statistics Simulations, **vol.** 15, pp. 74-81.

[85] Mittniks, S. and Rachevs, S.T. (1997). *Modeling Financial Assets with Alternative Stable Models.* Wiley, New York.

[86] Mittnik, S., and Rachev, S.T. (1993a). *Modeling asset returns with alternative stable models.* Econometric Reviews, **vol.** 12, pp. 261-330.

[87] Mittnik, S., and Rachev, S.T. (1993b). *Reply to comments on "Modeling asset returns with alternative stable models" and some extensions.* Econometric Reviews, **vol.** 12, pp. 347-389.

[88] Mittnik, S., Paolella, M.S., Rachev, S.T. (2000). *Diagnosing and treating the fat tails in financial returns data.* Journal of Empirical Finance, **vol.** 7, pp. 389-416

[89] Mittnik, S., Paolella, M.S., Rachev, S.T. (2002). *Stationary of stable power-GARCH processes.* Journal of Econometrics, **vol.** 106, pp. 97-107.

[90] Mokkadem, A. (1988). *Mixing Properties of ARMA Processes.* Stochastic Processes and their Applications, **vol.** 29, pp. 309-315.

[91] Painter, S. (1995). *Random fractal models of heterogeneity: the Lévy-stable approach.* Math. Geol., **vol.** 27, pp. 813-830.

[92] Painter, S. (1998). *Long-range dependence in the subsurface: Empirical evidence and simulation methods.* Invited paper at the American Geophysical Union, Fall Meeting.

[93] Peligrad, M. and Sang, H. (2012). *Asymptotic properties of self-normalized linear processes with long memory.* Econometric Theory, **vol.** 28, pp. 1-22.

[94] Phillips, P.C.B. and Solo, V. (1992). *Asymptotics for linear Processes.* The Annals of Statistics, **vol.** 20, pp. 971-1001.

[95] Philipp, W. (1986). *.Invariance principles for independent and weakly dependent random variables. In Dependence in Probability and Statistics: A Survey of Recent Results (E. Eberlein and M. S. Taqqu, eds.)* Birkhäuser, Boston.

[96] Philipp, W. and Stout, W. (1975). *Almost sure invariance principles for partial sums of weakly dependent random variables.* Mem. Amer. Math. Soc. 2.

[97] Poisson, S.D. (1837). *Récherches sur la probabilité des jugements en matiére criminelle et en matiére civile.*

[98] Racheva, B. and Samorodnitsky, G. (2003). *Long range dependence in heavy tailed stochastic processes, in Handbook of Heavy Tailed Distributions in Finance.* (edited by Rachev, S. (2003)). Elsevier: Amsterdam.

[99] Rachev, S., Menn, C. and Fabozzi, F. (2005). *Fat-Tailed and Skewed Asset Return Distributions.* John Wiley & Sons. Inc: Hoboken.,**vol.** 12, pp. 261-330.

[100] Resnick, S.I. (1986). *Point processes, regular variation and weak convergence.* Adv. Appl. Probab., **vol.** 18, pp. 66138.

[101] Rio, E. (1995). *A maximal inequality and dependent Marcinkiewicz-Zygmund strong laws.* Ann. Probab., **vol.** 2(23), pp. 918-937.

[102] Robinson, P. (ed.) (2003). *Time Series with Long Memory. Advanced Texts in Econometrics.* Oxford University Press.

[103] Robbins, H. and Monro, S. (1951). *A stochastic approximation method.* Ann. Math. statist., **vol.** 22, pp. 400-407.

[104] Rosenblatt, M. (1961). *Independence and dependence.* Proc. 4th Berkeley Symp. Math. Statist. Probab., pp. 411-443.

[105] Rosenblatt, M. (1984). *Stochastic processes with short-range and long-range dependence.* in Statistics: An Appraisal, (H. David and H. David, eds.), pp. 509-520, Iowa State University Press.

[106] Samorodnitsky, G. (2006). *Long memory and self-similar processes.* Annales de la Faculté des Sciences de Toulouse, **vol.** 15, pp. 107-123.

[107] Samorodnitsky, G. (2002). *Long Range Dependence, Heavy Tails and Rare Events.* MaPhySto, Centre for Mathematical Physics and Stochastics, Aarhus. Lecture Notes.

[108] Saridis, G.N. (1974). *Stochastic Approximation Methods for Identification and Control A Survey.* IEEE-AC, **vol.** 19, No 6.

[109] Shao, Q.M. (1993). *Almost sure invariance principles for mixing sequences of random variables.* Stochastic Process. Appl., **vol.** 48, pp. 319-334.

[110] Skorokhod, A.V. (1957). *Limit Theorems for Stochastic Processes With Independent Increments.* Theory Probab. Appl., **vol.** 2, pp. 145-177.

[111] Skorokhod, A.V. (1956). *Limit theorems for stochastic Processes.* Theory Probab. Appl., **vol.** 1, pp. 261-290.

[112] Soderstrom, T. and Stoica, P. (1989). *System Identification.* Prentice Hall.

[113] Solo, V. and Kong, X. (1995). *Adaptive Signal Processing Algorithms: Stability and Performance.* Englewood Cliffs, NJ: Prentice-Hall.

[114] Stout, W.F. (1974). *Almost Sure Convergence.* Academic Press, New York.

[115] Surgailis, D. (1982). *Zones of attraction of self-similar multiple integrals.* Lithuanian Math. J., **vol.** 22, pp. 327-340.

[116] Surgailis, D. (2004). *Stable limits of sums of bounded functions of long-memory moving averages with finite variance.* Bernoulli, **vol.** 10, pp. 327-355.

[117] Tadić, V.B. (2004). *On the Almost Sure Rate of Convergence of Linear Stochastic Approximation Algorithms.* IEEE Trans. Inform. Theory, **vol**. 50, No. 2, pp. 401-409.

[118] Taqqu, M.S. (1979). *Convergence of integrated processes of arbitrary Hermite rank.* Z. Wahrscheinlichkeitstheorie Verw. Geb., **vol.** 50, pp. 53-83.

[119] Taqqu, M. (1986). *A bibliographical guide to self-similar processes and long-range dependence.* in Dependence in Probability and Statistics, (E. Eberlein and M. Taqqu, eds.), pp. 137-162, Boston: Birkhäuser.

[120] Varotsos, C. and Kirk-Davidoff, D. (2006). *Long-memory processes in global ozone and temperature variations at the region* 60^0 *S*-60^0 *N.* Atmospheric Chemistry and Physics, **vol.** 6, pp. 4093-4100.

[121] Thanh, L.V., Yin, G. and Wang, L.Y. (2011). *State observers with random sampling times and convergence analysis of double-indexed and randomly-weighted sums of mixing processes.* SIAM J. Control Optim., **vol.** 49, No. 1, pp. 106-124.

[122] Vaiciulis, M. (2003). *Convergence of sums of Appell polynomials with infinite variance.* Lithuanian Math. J., **vol.** 43, pp. 80-98.

[123] Varotsos, C. and Kirk-Davidoff, D. (2006). *Long-memory processes in global ozone and temperature variations at the region* 60^0 *S*-60^0 *N.* Atmospheric Chemistry and Physics, **vol.** 6, pp. 4093-4100.

[124] de Vries, C.G. (1991). *On the relation between GARCH and stable processes.* Journal of Econometrics, **vol.** 48, pp. 313-324.

[125] Walk, H. and Zsidó, L. (1989). *Convergence of Robbins-Monro method for linear problems in banach space.* J. Math. Anal. Applic., **vol**. 139, pp. 152-177.

[126] Withers, C.S. (1981). *Conditions for linear processes to be strong-mixing.* Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete [Became: @J(ProbTher)], **vol.** 57, pp. 477-480.

[127] Widrow, B., Glover, J.R., McCool, J., Kaunitz, J., Williams, C.S., Hearn, R.H., Zeidler, J.R., Dong, E., Goodlin, R.C. (1975). *Adaptive Noise Cancelling : Principles and Applications.* Proc. IEEE, **vol.** 63, No 12.

[128] Wu, W.B. and Min, W. (2005). *On linear processes with dependent innovations.* Stochastic Processes and their Applications, **vol.** 115, pp. 939-958.

[129] Wu, W. B., Huang, Y. and Zheng, W. (2010). *Covariances Estimation for Long-Memory Processes.* Adv. Appl. Prob., **vol.** 42, pp. 137-157.

[130] Wu, W.B. (2007). *Strong Invariance Principles for Dependent Random Variables.* The Annals of Probability, **vol.** 35, pp. 2294-2320.

[131] Yaglom, A. (1955). *Correlation theory of processes with stationary random increments of order n.* Matematica Sbornik, **vol.** 37, pp. 141-196. (English translation in American Mathematical Society of Translations Series 2, **vol.** 8, pp. 87-141, 1958).

[132] Yin, G. (1992). *Asymptotic Optimal Rate of Convergence for an Adaptive Estimation Procedure.* Stochastic Theory and Adaptive Control, Lecture Notes in Control and Information Sciences, **vol.** 184, pp. 480-489.

[133] Zhanga, R., Sinb, C.Y., and Lingc, S. (2015). *On functional limits of short- and long-memory linear processes with GARCH(1,1) noises.* Stochastic Processes and their Applications, **vol.** 125, pp. 482-512.