# Strong Laws of Large Numbers Under Weak Assumptions with Application

Brett Ninness*

**Abstract**

The employment of 'Strong Laws of Large Numbers' is instrumental to the analysis of system estimation and identification strategies. However, the vast bulk of such laws, as presented in the wider literature, assume independence or at least uncorrelatedness of random components and these assumptions are quite restrictive from an engineering point of view. By way of contrast, this paper shows how to establish strong laws for possibly non-stationary random processes with very general dependence structure. Brief examples are provided that illustrate the utility of the Strong Law of Large Numbers presented.

Technical Report EE9881, Department of Electrical and Computer Engineering, University of Newcastle,AUSTRALIA

## 1 Introduction

It is common in the analysis of system identification and parameter estimation algorithms to consider their performance as the amount $N$ of available data samples increases [19, 28]. The goal in these instances is to prove convergence of estimates to values which (provided the parameterised model class is rich enough to describe the true dynamics) are the true parameters. Underlying this is a presumption that if an estimation algorithm is effective, then given an infinite amount of data, it should be able to extract parameter estimates perfectly, even if measurement corruptions exist; although it should be acknowledged that there are alternative points of view on this matter [15, 29].

In proving these sorts of results, the key step is the employment of a 'Strong Law of Large Numbers' (SLLN), by which is meant a law that a sum of random quantities converges to a known value, usually zero. For instance, considering the simplest possible example for motivational

*This work was supported by the Australian Research Council and the Centre for Integrated Dynamics and Control. This author is with the Department of Electrical and Computer Engineering, University of Newcastle, Australia and can be contacted at email:brett@ee.newcastle.edu.au or FAX: +61 2 49 21 69 93

purposes, suppose that it is necessary to estimate a constant scalar quantity $\theta_\circ$ on the basis of $N$ measurements $\{y_1, y_2, \cdots, y_N\}$ that are corrupted by a sequence of random variables $\{\nu_t\}$ as

$$y_t = \theta_\circ + \nu_t.$$

The least-squares estimate $\widehat{\theta}_N$ of $\theta_\circ$ would then be given as

$$\widehat{\theta}_N = \arg\min_{\theta \in \mathbf{R}} \sum_{t=1}^{N} (y_t - \theta)^2 = \frac{1}{N} \sum_{t=1}^{N} y_t$$

in which case the estimation error $\widetilde{\theta}_N \triangleq \widehat{\theta}_N - \theta_\circ$ is

$$\widetilde{\theta}_N = \frac{1}{N} \sum_{t=1}^{N} \nu_t.$$

Now, it is widely available [5, page 103] that if $\{\nu_t\}$ is an uncorrelated (white) sequence of random variables, with each $\nu_t$ defined on a probability space $\{\Omega, \mathcal{F}, \mathbf{P}\}$ such that $\mathbf{E}\{\nu_t^2\} < \infty$, then

$$\frac{1}{N} \sum_{t=1}^{N} \nu_t \xrightarrow{a.s.} 0 \qquad \text{as } N \to \infty \tag{1}$$

where the $a.s.$ (almost sure) epithet means that the above convergence, if it fails, does so only on a subset $\Omega' \subset \Omega$ such that $\mathbf{P}(\Omega') = 0$. Such a result as (1) is called a 'Strong Law of Large Numbers', and it is clear that by employing it, then under the afore-mentioned stochastic assumptions on $\{\nu_t\}$, it may be concluded that the estimation error $\widetilde{\theta}_N$ will tend almost surely to zero, in which case the estimation $\widehat{\theta}_N$ is termed 'strongly consistent'.

A natural question at this stage is that of rate of convergence. In one instance, this can be answered by imposing the further restriction that the $\{\nu_t\}$ are not only uncorrelated, but are in fact independent, in which case it is again widely available (see, for example, [5, Thm. 5.4.1]) that for any $\alpha > 0$

$$\frac{1}{N^{1/2+\alpha}} \sum_{t=1}^{N} \nu_t \xrightarrow{a.s.} 0 \qquad \text{as } N \to \infty \tag{2}$$

which gauges the convergence rate $\widetilde{\theta}_N \to 0$ as almost $1/\sqrt{N}$. Under an additional assumption of identical distribution of the $\{\nu_t\}$, the famous Law of the Iterated Logarithm [25, Thm. 2.8.2], [8, Thm. 12.5.1] indicates that $\sqrt{(\log \log N)/N}$ over-bounds the actual convergence rate of (1).

Unfortunately, from the perspective of utility in engineering relevant applications, assumptions of independence or of identical distribution of the measurement error process $\{\nu_t\}$ components are overly restrictive. Instead, laws of large numbers are required that eschew these restrictions in favour of ones that are more likely to allow accurate modelling of the error processes that are typically observed in practice.

A difficulty then arises, in that the required SLLN results do not seem to be readily available in the literature. For example, with reference to the suite of monographs in the probability theory

literature, the majority [5, 8, 3, 30, 27, 1, 24, 9] do not move beyond the uncorrelated case in the presentation of laws of large numbers, while those that do progress to the dependent data case [30, 25], do so either by moving to a martingale formulation, or by using the idea of 'mixing'. To engineers without specialist probability theory backgrounds, it may be quite difficult to assess the physical implications of these frameworks which depend (respectively) on bounds on conditional expectations or conditional probabilities. In particular, from an engineering perspective, just how much milder is it to assume that a sequence is a martingale difference rather than white noise?

The purpose of this paper is to provide some redress to this situation by establishing that the strong law of large numbers (2) still holds for a very general class of dependent, possibly non-stationary sequences of random variables which can be very simply characterised in a manner that makes clear the restrictions being imposed on the correlation structure of the process. Underlying this paper is an assumption that such assumptions on dependence are the most natural and transparent from an engineering perspective.

For example, the work here will establish that any process $\{\nu_t\}$ for which the $N \times N$ matrix $[R_t]_{m,n} = \mathbf{E}\{\nu_{t+m}\nu_{t+n}\}$ satisfies $N^{-1}\|R_t\|_2 \leq C < \infty$ for some $C$ independent of $N$ or $t$ obeys (2). A particular instance of such processes is shown to be stationary processes with bounded spectral density.

Prior to these results, to the authors knowledge the most general preceding them was that employed by Ljung in [20, 18] and attributed to Cramér and Leadbetter (who established it in continuous time in their 1967 monograph [6]) in which (1) is asserted to hold provided that

$$|\mathbf{E}\{\nu_t\nu_s\}| \leq C\frac{t^p + s^p}{1 + |t-s|^q}, \qquad 0 \leq 2p < q < 1.$$

However, using the results of this paper, under the same assumption the stronger result (2) is established for any $\alpha > (1+p-q)/2$, with the dividend being that an estimate of the convergence rate of (the $1/N$ normalised partial sum) (1) is then established as being at least $N^{(p-q)/2}$.

Other general results [22, 26, 21] that employ the same 'maximal-inequality' tools used here have been provided by anonymous reviewers, and their relationship to this paper will be commented on presently.

# 2 Main Result

Without further preliminaries, the main result of the paper is the following strong law of large numbers.

**Theorem 2.1.** *Suppose $\{\nu_t\}$ is a sequence of random variables, not necessarily zero mean, and with arbitrary correlation structure (not necessarily stationary) that is characterised by the existence of a $C < \infty$, $1 < \beta < \infty$ such that*

$$\sum_{t=1}^{N}\sum_{s=1}^{N} \mathbf{E}\{\nu_t\nu_s\} \leq CN^{\beta}.$$

*Then for any* $\alpha > \beta/2$

$$\frac{1}{N^\alpha}\sum_{t=1}^{N}\nu_t \xrightarrow{a.s.} 0 \qquad \textit{as } N \to \infty.$$

*Proof.* Define

$$S_N \triangleq \sum_{t=1}^{N}\nu_t$$

and, for any $N$, choose an integer $M$ such that $2^M \le N \le 2^{M+1}$. Then

$$\frac{1}{N^\alpha}|S_N| \le \frac{1}{2^{M\alpha}}\max_{0\le k\le 2^{M+1}}|S_k|$$

and, by the assumptions of the Theorem

$$\mathbf{E}\left\{\left|\sum_{t=1}^{2^{M+1}}\nu_t\right|^2\right\} = \sum_{t=1}^{2^{M+1}}\sum_{s=1}^{2^{M+1}}\mathbf{E}\left\{\nu_t\nu_s\right\} \le C2^{(M+1)\beta}.$$

Therefore, by employing Lemma A.1 with $u_t = 1$ and for arbitrary $\epsilon > 0$

$$\mathbf{P}\left\{\max_{1\le k\le 2^{M+1}}|S_k| \ge \epsilon 2^{M\alpha}\right\} \;\le\; \frac{C2^{(M+1)\beta}}{2^{2\alpha M}} = \frac{2^\beta C}{2^{(2\alpha-\beta)M}}.$$

Consequently, for any $\alpha > \beta/2$

$$\sum_{M=1}^{\infty}\mathbf{P}\left\{\max_{0\le k\le 2^{M+1}}|S_k| \ge \epsilon 2^{\alpha M}\right\} \le 2^\beta C\sum_{M=1}^{\infty}\frac{1}{2^{(2\alpha-\beta)M}} < \infty$$

so that by the Borel-Cantelli Lemma [5, 3, 30]

$$\limsup_{N\to\infty}\frac{1}{N^\alpha}|S_N| \le \limsup_{M\to\infty}\frac{1}{2^{\alpha M}}\max_{0\le k\le 2^{M+1}}|S_k| = 0$$

with probability one. □

A key feature of this result is that the sum of $\nu_t$'s is normalised by a factor $1/N^\alpha$ which, depending on $1 < \beta < 2\alpha$, may be greater than the factor $1/N$ which appears in most SLLN's available in the literature; for example, the Cramér and Leadbetter result [6] previously mentioned. The significance of this, as will be illustrated for estimation algorithms in §4, is that in addition to being able to verify convergence itself, a certain flexibility is available which allows upper bounds on the rate of convergence to also be established.

Thanks to the work of anonymous reviewers, the results most closely related to Theorem 2.1 appear to be contained in Theorem 6 of [22] and Theorems 3.7.2, 3.7.6 of [26]. However the key differences between these works and Theorem 2.1 is that either (Theorem 3.7.2 of [26]) the normalisation of the sum is fixed at $1/N$ (instead of being a variable amount $N^\alpha$ that depends

on the memory of $\{\nu_t\}$), or (Theorem 6 of [22], Theorem 3.7.6 of [26]) the conditions on the memory of $\{\nu_t\}$ involve a bound on $\mathbf{E}\{|\sum_{t=1}^{N}\nu_t|^\gamma\}$ with $\gamma > 2$ while Theorem 2.1 allows the case of $\gamma = 2$ which, as the following section will illustrate, is far easier to work with.

As a final general remark, notice that although Theorem 2.1 allows (via the possibility of $\beta > 2$) normalisation in the partial sum by a factor $1/N^\alpha$ with $\alpha > 1$, such a situation is not common in engineering applications.

# 3 Specific Cases

Having provided the convergence result of theorem 2.1, the paper now goes on to enumerate specific engineering relevant classes of stochastic processes $\{\nu_t\}$ for which theorem 2.1 is applicable. These cases are organised as corollaries to the main theorem, beginning with that of wide-sense stationary stochastic processes.

**Corollary 3.1.** *Suppose that $\{\nu_t\}$ is a zero mean wide-sense stationary stochastic process with associated spectral density $f_\nu(\omega)$ bounded as $\|f_\nu\|_\infty \leq C < \infty$. Then for any $\alpha > 0$*

$$\frac{1}{N^{1/2+\alpha}}\sum_{t=1}^{N}\nu_t \xrightarrow{a.s.} 0 \qquad \text{as } N \to \infty. \tag{3}$$

*Proof.* Under the assumptions of the corollary

$$\begin{aligned}\sum_{t=1}^{N}\sum_{s=1}^{N}\mathbf{E}\{\nu_t\nu_s\} &= \sum_{t=1}^{N}\sum_{s=1}^{N}\frac{1}{2\pi}\int_{-\pi}^{\pi} f_\nu(\omega)\, e^{j(t-s)\omega}\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f_\nu(\omega)\left|\sum_{t=1}^{N} e^{jt\omega}\right|^2 \mathrm{d}\omega \\ &\leq C\sum_{t=1}^{N}\sum_{s=1}^{N}\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j(t-s)\omega}\,\mathrm{d}\omega = CN.\end{aligned}$$

Applying theorem 2.1 with $\beta$ arbitrarily close to $1$ then provides the result. □

To the author's knowledge, the only other generally available references providing SLLN's for wide-sense stationary stochastic processes are [7, 23, 13] for which the weaker result (1) is presented. However, note that under stronger conditions than imposed in Corollary 3.1 which involve smoothness restrictions on the spectral factor of $f_\nu(\omega)$ (or, what is equivalent, assumptions on rates of decay of Fourier co-efficients of the spectral factor), the result (3) can be inferred from the so called 'Beveridge–Nelson' decomposition techniques of [23].

Moving on to the non-stationary case:

**Corollary 3.2.** *Suppose that $\{\nu_t\}$ is a possibly non-stationary stochastic process satisfying, for some $C < \infty$, the (relaxed) 'Cramér-Leadbetter condition'*

$$|\mathbf{E}\{\nu_t\nu_s\}| \leq C\frac{t^p + s^p}{1 + |t-s|^q}, \qquad ; p < q < 1.$$

*Then for any* $\epsilon > 0$

$$\frac{1}{N^{1-(q-p)/2+\epsilon}} \sum_{t=1}^{N} \nu_t \xrightarrow{a.s.} 0 \qquad \textit{as } N \to \infty.$$

*Proof.*

$$\begin{aligned}
\sum_{t=1}^{N}\sum_{s=1}^{N} \frac{t^p + s^p}{1+|t-s|^q} &= 2\sum_{t=1}^{N} t^p + 2\sum_{t=1}^{N}\sum_{s=1}^{t-1} \frac{t^p + s^p}{1+(t-s)^q} \\
&\le 2\int_0^N t^p \,\mathrm{d}t + 2\sum_{t=1}^{N} t^p \sum_{s=1}^{t-1} \frac{1}{1+(t-s)^q} + 2\sum_{t=1}^{N}\sum_{s=1}^{t-1} \frac{s^p}{1+(t-s)^q} \\
&\le \frac{2N^{p+1}}{p+1} + 4\sum_{t=1}^{N} t^p \sum_{s=1}^{t-1} \frac{1}{1+(t-s)^q}.
\end{aligned}$$

However

$$\begin{aligned}
\sum_{t=1}^{N} t^p \sum_{s=1}^{t-1} \frac{1}{1+(t-s)^q} &\le \int_0^N t^p \left( \int_0^t \frac{1}{1+(t-s)^q} \,\mathrm{d}s \right) \mathrm{d}t \\
&\le \int_0^N t^p \left( 1 + \int_1^t \frac{1}{x^q} \,\mathrm{d}x \right) \mathrm{d}t \\
&= \frac{1}{1-q} \int_0^N (t^{p-q+1} - qt^p) \,\mathrm{d}t < CN^{p-q+2}.
\end{aligned}$$

Therefore, there exists a $C < \infty$ such that

$$\sum_{t=1}^{N}\sum_{s=1}^{N} \mathbf{E}\{\nu_t \nu_s\} \le CN^{2+(p-q)}$$

and the corollary then follows by direct application of Theorem 2.1. □

The '(relaxed) Cramér-Leadbetter' condition quoted in Corollary 3.2 earns its epithet from being somewhat milder than that originally formulated in [6] (and used in various works such as [20, 10]) wherein the restriction $0 \le 2p < q < 1$ is imposed.

In system identification applications, in addition to these results on the behaviour of sample means, results on the convergence of sample covariances are also useful. However, these results are more difficult to derive (or to find in general form in the literature) than results on sample means. The following corollary to Theorem 2.1 is provided in consideration of this, but because of particular difficulties with the higher order moments involved, it imposes more restrictive assumptions than considered in the previous results of this section.

**Corollary 3.3.** *Suppose that* $\{\nu_t\}$ *and* $\{z_t\}$ *are zero mean wide-sense stationary stochastic processes formed as*

$$\nu_t = \sum_{k=0}^{\infty} h_k e_{t-k}, \qquad z_t = \sum_{k=0}^{\infty} g_k e_{t-k}$$

*where* $\{e_t\}$ *is a zero mean i.i.d. process with* $\boldsymbol{E}\{e_t^4\} < \infty$ *and* $\sum_{k=0}^{\infty} h_k^2 < \infty$, $\sum_{k=0}^{\infty} g_k^2 < \infty$. *Then for any* $\alpha > 0$

$$\frac{1}{N^{1/2+\alpha}} \sum_{t=1}^{N} \left[\nu_t z_{t+\tau} - \boldsymbol{E}\left\{\nu_t z_{t+\tau}\right\}\right] \xrightarrow{a.s.} 0 \qquad \textit{as } N \to \infty. \tag{4}$$

*Proof.* By repeatedly using the independence and zero mean assumptions:

$$\mathbf{E}\left\{\left(\sum_{t=1}^{N} \nu_t z_{t+\tau} - \mathbf{E}\left\{\nu_t z_{t+\tau}\right\}\right)^2\right\} =$$

$$\begin{aligned}
&\sum_{t=1}^{N}\sum_{s=1}^{N}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\sum_{\ell=0}^{\infty}\sum_{k=0}^{\infty} h_n g_m h_k g_\ell \times \\
&\left[\mathbf{E}\left\{e_{t-k}e_{t+\tau-\ell}e_{s-n}e_{s+\tau-m}\right\} - \mathbf{E}\left\{e_{t-k}e_{t+\tau-\ell}\right\}\mathbf{E}\left\{e_{s-n}e_{s+\tau-m}\right\}\right] \\
= &\sum_{t=1}^{N}\sum_{s=1}^{N}\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} h_n g_{n+\tau} h_k g_{k+\tau} \left[\mathbf{E}\left\{e_{t-k}^2 e_{s-n}^2\right\} - \mathbf{E}\left\{e_{t-k}^2\right\}\mathbf{E}\left\{e_{s-n}^2\right\}\right] \\
= &\sum_{t=1}^{N}\sum_{s=1}^{N}\sum_{n=0}^{\infty} h_n g_{n+\tau} h_{n+t-s} g_{n+t+\tau-s} \left[\mathbf{E}\left\{e_{s-n}^4\right\} - \left(\mathbf{E}\left\{e_{s-n}^2\right\}\right)^2\right] \\
\leq &\; C\sum_{t=1}^{N}\sum_{s=1}^{N}\sum_{n=0}^{\infty} h_n g_{n+\tau} h_{n+t-s} g_{n+t+\tau-s}
\end{aligned}$$

where $C \leq \mathbf{E}\{e_{s-n}^4\} - (\mathbf{E}\left\{e_{s-n}^2\right\})^2 < \infty$. Now, by the assumption that $\sum h_k^2 < \infty$, $\sum g_k^2 < \infty$ and by the Cauchy–Schwarz inequality

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} |h_n g_{n+\tau} h_{m+t-s} g_{m+t+\tau-s}| \;\leq\; \left(\sum_{n=0}^{\infty} |h_n g_{n+\tau}|\right)^2 \leq \left(\sum_{n=0}^{\infty} h_n^2\right)\left(\sum_{n=0}^{\infty} g_n^2\right) < \infty.$$

Because of this absolute summability, exchange of integration and summation is permitted to obtain (assuming $h_n = 0$ for $n < 0$)

$$\sum_{t=1}^{N}\sum_{s=1}^{N}\sum_{n=0}^{\infty} h_n g_{n+\tau} h_{n+t-s} g_{n+t+\tau-s} =$$

$$\begin{aligned}
&\frac{1}{2\pi}\sum_{t=1}^{N}\sum_{s=1}^{N}\int_{-\pi}^{\pi}\sum_{m=-\infty}^{\infty}\sum_{n=0}^{\infty} h_n g_{n+\tau} h_{m+t-s} g_{m+t+\tau-s} e^{j\omega(n-m)}\,\mathrm{d}\omega \\
= &\;\frac{1}{2\pi}\sum_{t=1}^{N}\sum_{s=1}^{N}\int_{-\pi}^{\pi} e^{j\omega(t-s)}\left(\sum_{n=0}^{\infty} h_n g_{n+\tau} e^{j\omega n}\right) \times
\end{aligned}$$

$$\begin{aligned} & \left( \sum_{m=-\infty}^{\infty} h_{m+t-s} g_{m+t+\tau-s} e^{-j\omega(m+t-s)} \right) \mathrm{d}\omega \\ = \; & \frac{1}{2\pi} \sum_{t=1}^{N} \sum_{s=1}^{N} \int_{-\pi}^{\pi} e^{j\omega(t-s)} |F_\tau(\omega)|^2 \, \mathrm{d}\omega \\ \leq \; & N \max_{\omega} |F_\tau(\omega)|^2 \end{aligned}$$

where $F_\tau(\omega) \triangleq \sum_{n=0}^{\infty} h_n g_{n+\tau} e^{j\omega\tau}$, and again by the Cauchy–Schwarz inequality

$$|F_\tau(\omega)|^2 \leq \left( \sum_{n=0}^{\infty} h_n^2 \right) \left( \sum_{n=0}^{\infty} g_n^2 \right) < \infty.$$

Therefore, there exists a $C < \infty$ such that

$$\mathsf{E} \left\{ \left( \sum_{t=1}^{N} \nu_t z_{t+\tau} - \mathsf{E}\left\{ \nu_t z_{t+\tau} \right\} \right)^2 \right\} < CN$$

so that the result of the corollary follows by direct application of Theorem 2.1. □

In the literature, the result closest to this corollary that is known to the author is provided in [23] where a similar construction of $\{\nu_t\}$ is supposed (cross-correlation involving $z_t$ is not considered), save that only $\mathsf{E}\{e_t^2\} < \infty$ is required, but a stronger assumption of $\sum k h_k^2 < \infty$ is imposed. As well, in that work (which, despite its breadth, does not appear to be well known in engineering circles) the methods involve ideas of martingales adapted to filtrations, which may make the derivations less widely accessible than the one presented here that only requires expectations to be bounded by using the Cauchy–Schwarz inequality. Results on convergence of sample covariances are also provided in [11, Theorem 6.3.5], but deal only with convergence in probability, not almost-sure convergence.

# 4 Applications

Having provided the convergence results that were the main purpose of the paper, the remainder of the work is devoted to illustrations of how they may be gainfully employed in engineering relevant situations, all of which pertain to estimation problems. The consistency results to be presented are not new, but as far as the author is aware, the almost-sure convergence rate results have not appeared before. In any event, the purpose of this section is not to establish new results, but to illustrate the utility of Theorem 2.1 and its corollaries.

## 4.1 Estimation using Orthonormal Bases

As a first simple example, suppose that for some $n$ dimensional parameter vector $\theta_\circ \in \mathbf{R}^n$, the relationship between an observed input sequence $\{u_t\}$ and output sequence $\{y_t\}$ obeys

$$y_t = \phi_t^T \theta_\circ + \nu_t \tag{5}$$

where $\{\nu_t\}$ is stationary stochastic process with bounded spectral density $f_\nu(\omega) < C < \infty$ which is uncorrelated with any elements in the regression vector $\phi_t$.

This uncorrelatedness occurs (for example) when the elements of $\phi_t$ depend only upon the input sequence $\{u_t\}$ and, in turn, this can arise when (5) corresponds to expressing the input-output dynamics as a linear combination of filters $\mathcal{B}_1(q), \cdots, \mathcal{B}_n(q)$:

$$y_t = \sum_{k=1}^{n} \theta_k \mathcal{B}_k(q) u_t = \phi_t^T \theta \tag{6}$$

where

$$\phi^T \triangleq [\mathcal{B}_1(q)u_t, \mathcal{B}_2(q)u_t, \cdots, \mathcal{B}_n(q)u_t], \qquad \theta^T \triangleq [\theta_1, \theta_2, \cdots, \theta_n] \in \mathbf{R}^n.$$

In particular, $\mathcal{B}_k(q) = q^{-k}$ corresponds to the common case of FIR modelling, but all that will be assumed here is that all the $\{\mathcal{B}_k(q)\}$ are strictly stable.

In engineering settings, it is common to require an estimate $\widehat{\theta}_N$ of $\theta_\circ$ that is based on $N$ observations $\{u_1, \cdots, u_N\}$, $\{y_1, \cdots, y_N\}$ of input and output. In turn, this is commonly achieved by choosing $\widehat{\theta}_N$ so as to minimise the sum of the squared errors between the observation $\{y_1, \cdots, y_N\}$ and the model $\phi_t^T\theta$, which is well known [19, 28] to be given as the solution to

$$\left(\frac{1}{N}\sum_{t=1}^{N} \phi_t \phi_t^T\right) \widehat{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} \phi_t y_t. \tag{7}$$

Then defining the parameter estimation error $\widetilde{\theta}_N \triangleq \widehat{\theta}_N - \theta_\circ$ allows (7) to be re-expressed as

$$\left(\frac{1}{N}\sum_{t=1}^{N} \phi_t \phi_t^T\right) \widetilde{\theta}_N = \frac{1}{N}\sum_{t=1}^{N} \phi_t \nu_t. \tag{8}$$

Suppose further that $\{u_t\}$ is quasi-stationary in the sense used by Ljung [19] (amenable to Wiener's Generalised Harmonic Analysis [4]) so that the following limit exists

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} u_t u_{t+\tau} = R_u(\tau) \tag{9}$$

(for example, by Corollary 3.3 this limit will exist almost surely if $\{u_t\}$ is generated as appropriately filtered white noise) where $R_u(\tau)$ is such that the associated spectral density $f_u(\omega) = \sum_\tau R_u(\tau)e^{-j\omega\tau}$ is bounded away from zero. Then

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} \phi_t \phi_t^T = P > 0$$

which implies that for some $C' < \infty$ the solution to (8) must satisfy (denoting $\theta(m)$ and $\phi_t(m)$ as the $m$'th components of those vectors)

$$\left|\widetilde{\theta}_N(m)\right| \leq \frac{C'}{N}\left|\sum_{t=1}^{N} \phi_t(m)\nu_t\right|. \tag{10}$$

Furthermore, if the input $\{u_t\}$ is also assumed to be bounded in magnitude then $|\phi_t(m)| < C''$ for some $C'' < \infty$ so that

$$\begin{aligned}\sum_{t=1}^{N}\sum_{s=1}^{N}\phi_t(m)\phi_s(m)\mathbf{E}\{\nu_t\nu_s\} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f_\nu(\omega)\mathbf{E}\left\{\left|\sum_{t=1}^{N}\phi_t(m)e^{j\omega t}\right|^2\right\}\mathrm{d}\omega \\ &\leq C\sum_{t=1}^{N}\phi_t^2(m) \leq C(C'')^2N.\end{aligned}$$

Therefore, by Theorem 2.1, for any $\epsilon > 0$

$$\frac{1}{N^{1/2+\epsilon}}\sum_{t=1}^{N}\phi_t\nu_t \xrightarrow{a.s.} 0, \qquad \text{as } N \to \infty. \tag{11}$$

Combining this with (10) then provides the conclusion that for any $\alpha < 1/2$

$$N^\alpha(\widehat{\theta}_N - \theta_\circ) \xrightarrow{a.s.} 0, \qquad \text{as } N \to \infty \tag{12}$$

and hence the least-squares estimate defined by (7) is strongly consistent.

The virtue of Theorem 2.1 is that it allowed this result to be derived very directly and under quite mild assumptions on disturbance and inputs. Furthermore, note that a direct consequence of Theorem 2.1 adapting the sum normalisation $1/N^\alpha$ to the dependence bound $N^\beta$, is that the convergence (12) then involves an $N^\alpha$ term, and this in turn makes available the rate-of-convergence bound

$$\widehat{\theta}_N - \theta_\circ = o(N^{-\alpha}), \qquad \alpha < 1/2 \tag{13}$$

which applies with probability one.

The more common $1/N$-normalised SLLN results in the literature clearly can not provide this sort of rate-bound. This has led to the development of alternate rate bounding methods that involve first establishing asymptotic distributional results, but then the ensuing rate estimates apply only in mean square or in probability, but not with probability one [19, 14].

Finally, note that the ability to provide a rate bound such as (13) does not depend on the modelling of $\{\nu_t\}$ as a stationary process. However, under more relaxed assumptions, the rate of convergence may be slower.

For example, if $\{\nu_t\}$ is allowed to be non-stationary but with a dependence structure that obeys for some $C < \infty$ the (relaxed) 'Cramér-Leadbetter' condition

$$|\mathbf{E}\{\nu_t\nu_s\}| \leq C\frac{t^p + s^p}{1 + |t-s|^q}, \qquad ; p < q < 1.$$

then since $|\phi_t(m)| < C''$

$$|\mathbf{E}\{\phi_t(m)\nu_t\phi_t(m)\nu_s\}| \leq C(C'')^2\frac{t^p + s^p}{1 + |t-s|^q}, \qquad ; p < q < 1.$$

so that by Corollary 3.2, for any $\alpha > 0$

$$\frac{1}{N^{1-(q-p)/2+\alpha}} \sum_{t=1}^{N} \phi_t \nu_t \xrightarrow{a.s.} 0, \qquad \text{as } N \to \infty.$$

and hence strong consistency is again established

$$N^{(q-p)/2+\alpha}\, \widehat{\theta}_N \xrightarrow{a.s.} \theta_\circ, \qquad \text{as } N \to \infty$$

but this time, with a rate of convergence diminished from that stationary $\{\nu_t\}$ case. The point is that the relationship between convergence rate and assumed dependence structure for $\{\nu_t\}$ is clearly exposed (it is proportional to $q - p$), and this exposure is not the case when rates are inferred from asymptotic distributional results.

Readers seeking a more sophisticated discussion of the application of stochastic convergence results to estimation problems are referred to[12, 16] and their bibliographies.

## 4.2 ARMA Modelling of Time Series

As a final illustration of how Theorem 2.1 and its Corollaries may be employed, consider the problem of estimating an ARMA model

$$y_t = \left( \frac{1 + c_1 q^{-1} + \cdots + c_n q^{-n}}{1 + d_1 q^{-1} + \cdots + d_n q^{-n}} \right) e_t = \left( \frac{C_\theta(q)}{D_\theta(q)} \right) e_t = H_\theta(q) e_t \tag{14}$$

on the basis of observing an $N$-sample realisation $\{y_1, y_2, \cdots, y_N\}$ of a wide-sense stationary time series. It will be assumed that $\{y_t\}$ is generated as

$$y_t = e_t + \sum_{k=1}^{\infty} h_k^\circ e_{t-k} = H_\circ(q) e_t, \qquad \sum_{k=1}^{\infty} (h_k^\circ)^2 < \infty \tag{15}$$

where $\{e_t\}$ is a zero mean i.i.d. process with $\mathbf{E}\{e_t^2\} = \sigma^2$, $\mathbf{E}\{e_t^4\} < \infty$. As well, the $\theta$ subscripting notation in the model (14) refers to it being parameterised by a vector

$$\theta^T = [c_1, c_2, \cdots, c_n, d_1, d_2, \cdots, d_n].$$

Estimating a model (14) on the basis of the $N$ measurements $\{y_1, y_2, \cdots, y_N\}$ thus amounts to deriving an estimated vector, call it $\widehat{\theta}_N$. There are many ways that this might be approached, but ideas of Maximum Likelihood motivate the formation of the Wiener filter, based on the model, and generating the mean square optimal one step ahead predictor $\widehat{y}_t(\theta)$ of $y_t$ as [14, 19]

$$\widehat{y}_t(\theta) = [1 - H_\theta^{-1}(q)] y_t$$

and then taking $\widehat{\theta}_N$ as a minimiser

$$\widehat{\theta}_N = \arg\min_{\theta \in \Theta} V_N(\theta) \tag{16}$$

of the cost

$$V_N(\theta) \triangleq \frac{1}{N}\sum_{t=1}^{N} \varepsilon_t^2(\theta) \tag{17}$$

which involves the error between the predictor and observations as

$$\varepsilon_t(\theta) = y_t - \widehat{y}_t(\theta).$$

In (16), the set $\Theta \subset \mathbf{R}^{2n}$ is defined as the set of parameters for which $H_\theta(q)$ and $H_\theta^{-1}(q)$ possesses square summable impulse responses. Typically, this is maintained by ensuring that the poles and zeros of $H_\theta(q)$ are strictly inside the unit disk.

Now, $\varepsilon_t(\theta) = H_\theta^{-1}(q)H(q)e_t$, so that by the definition of $\Theta$, the conditions of Corollary 3.3 are satisfied for all $\theta \in \Theta$ and hence

$$\sup_{\theta\in\Theta} \left| \frac{1}{N}\sum_{t=1}^{N} \left(\varepsilon_t^2(\theta) - \mathbf{E}\left\{\varepsilon_t^2(\theta)\right\}\right) \right| \xrightarrow{a.s.} 0, \qquad \text{as } N \to \infty$$

Since this convergence is uniform in $\theta \in \Theta$, which is an open subset of $\mathbf{R}^{2n}$, it is possible to conclude that the minima of $V_N(\theta)$ and $\mathbf{E}\{V_N(\theta)\}$ asymptotically co-incide with probability one [4, 11]. That is,

$$\widehat{\theta}_N \xrightarrow{a.s.} \theta_\star \triangleq \arg\min_{\theta\in\Theta} \lim_{N\to\infty} \mathbf{E}\left\{V_N(\theta)\right\} = \arg\min_{\theta\in\Theta} \mathbf{E}\left\{\varepsilon_t^2(\theta)\right\}$$

as $N \to \infty$ where $\mathbf{E}\{\varepsilon_t^2(\theta)\}$ is expressible in terms of the spectral density of $\varepsilon_t(\theta)$ as

$$\mathbf{E}\left\{\varepsilon_t^2(\theta)\right\} = \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} \left|\frac{H_\circ(e^{j\omega})}{H_\theta(e^{j\omega})}\right|^2 \mathrm{d}\omega. \tag{18}$$

Now, by Parseval's Theorem, this integral is equal to $\sum_{k\geq 0} c_k^2$ where the $c_k$ are the Fourier coefficients of $H_\circ(e^{j\omega})/H_\theta(j\omega)$, and by (14), (15) $c_0 = 1$ regardless of the choice of $\theta$. Therefore, if the model (14) is rich enough that there exists a $\theta_\circ$ such that $H_{\theta_\circ} = H_\circ$, then it is the unique minimiser ($c_k = 0$, $k \geq 1$) of (18) so that $\theta_\star = \theta_\circ$ and hence the strong consistency

$$\widehat{\theta}_N \xrightarrow{a.s.} \theta_\circ, \qquad \text{as } N \to \infty$$

is established. However, something more than just consistency is possible since the flexibility of Theorem 2.1 allows Corollary 3.3 to assert convergence for normalising factors converging to zero slower than $1/N$. This allows (as per § 4.1) an upper bound on the rate of convergence of $\widehat{\theta}_N$ to $\theta_\circ$ to be established as follows.

Firstly, using the notation $\cdot'$ to denote differentiation with respect to $\theta$, then since by the definition of $\widehat{\theta}_N$ the derivative $V_N'(\widehat{\theta}_N) = 0$, and by using the Mean Value Theorem, there exists a $\mu \in [0, 1]$ such that

$$V_N'(\theta_\circ) = V_N''(\zeta_N)(\theta_\circ - \widehat{\theta}_N) \tag{19}$$

where $\zeta_N = \mu\,\widehat{\theta}_N + (1-\mu)\theta_\circ$. Now, using the definition (17) it is straightforward to calculate that

$$V_N'(\theta) = -\frac{1}{N}\sum_{t=1}^{N} \varepsilon_t(\theta)\psi_t(\theta)$$

where $\psi_t(\theta)$ is the 'predictor gradient' defined by

$$\psi_t(\theta) \triangleq \frac{\mathrm{d}}{\mathrm{d}\theta}\widehat{y}_t(\theta) = H_\theta^{-1}(q)H_\theta'(q)\varepsilon_t(\theta) = H_\theta^{-2}(q)H_\theta'(q)H_\circ(q)e_t.$$

Therefore, since $\varepsilon_t(\theta) = H_\theta^{-1}H_\circ e_t$ and all the elements of $\psi_t(\theta)$ are formed (under the assumption on $\Theta$) as $\ell_2$ stably filtered versions of $\{e_t\}$, then Corollary 3.3 establishes that for any $\alpha < 1/2$

$$N^\alpha V_N'(\theta_\circ) = \frac{1}{N^{1-\alpha}}\sum_{t=1}^{N}\left[\mathbf{E}\left\{\varepsilon_t(\theta_\circ)\psi_t(\theta_\circ)\right\} - \varepsilon_t(\theta_\circ)\psi_t(\theta_\circ)\right] \xrightarrow{a.s.} 0 \qquad \text{as } N\to\infty.$$

But by the definition of $\theta_\circ = \theta_\star$

$$\mathbf{E}\left\{\varepsilon_t(\theta_\circ)\psi_t(\theta_\circ)\right\} = \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{E}\left\{V_N(\theta)\right\}\Big|_{\theta=\theta_\star} = 0$$

so that for any $\alpha < 1/2$

$$-N^\alpha V_N'(\theta_\circ) \xrightarrow{a.s.} 0 \qquad \text{as } N\to\infty. \tag{20}$$

Similarly,

$$V_N''(\zeta_N) = \frac{1}{N}\sum_{t=1}^{N}\psi_t(\zeta_N)\psi_t^T(\zeta_N) - \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t(\zeta_N)\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\psi_t(\theta)\Big|_{\theta=\zeta_N}\right)^T$$

so that again by virtue of all elements being $\ell_2$ stably filtered versions of $e_t$ and hence by Corollary 3.3

$$V_N''(\zeta_N) \xrightarrow{a.s.} P(\zeta_N) - Q(\zeta_N) \qquad \text{as } N\to\infty.$$

where

$$P(\zeta_N) \triangleq \mathbf{E}\left\{\psi_t(\zeta_N)\psi_t^T(\zeta_N)\right\}, \qquad Q(\zeta_N) \triangleq \mathbf{E}\left\{\varepsilon_t(\zeta_N)\psi_t'(\zeta_N)\right\}$$

However, since $\widehat{\theta}_N \xrightarrow{a.s.} \theta_\circ$ then $\widehat{\zeta}_N \xrightarrow{a.s.} \theta_\circ$ so that with probability one

$$\lim_{N\to\infty} Q(\zeta_N) = \mathbf{E}\left\{\varepsilon_t(\theta_0)\psi_t'(\zeta_N)\right\} = 0$$

where the equality to zero follows since $\varepsilon_t(\theta_0) = e_t$ which is zero mean and independent of $\psi_t'(\theta_\circ)$ which only contains terms involving $\{e_{t-1}, e_{t-2}, \cdots,\}$.

Furthermore, if $x, y \in \mathbf{R}^n$ are both such that $x^T x = y^T y = 1/2$ and $z$ is formed as $z = [x^T, y^T]^T$ then $z^T z = 1$ and hence by the spectral representation of $\mathbf{E}\{\psi_t(\theta_\circ)\psi_t^T(\theta_\circ)\}$

$$\begin{aligned} z^T P(\theta_\circ) z &= \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi} |C_{\theta_\circ}(e^{j\omega})|^{-2} \left|\sum_{k=1}^{n} x_k e^{j\omega k}\right|^2 + |D_{\theta_\circ}(e^{j\omega})|^{-2} \left|\sum_{k=1}^{n} y_k e^{j\omega k}\right|^2 \, \mathrm{d}\omega \\ &\geq \frac{\sigma^2}{2\pi}\min_{\omega} |C_{\theta_\circ}(e^{j\omega})|^{-2} \int_{-\pi}^{\pi} \left|\sum_{k=1}^{n} x_k e^{j\omega k}\right|^2 \, \mathrm{d}\omega + \\ &\quad \frac{\sigma^2}{2\pi}\min_{\omega} |D_{\theta_\circ}(e^{j\omega})|^{-2} \int_{-\pi}^{\pi} \left|\sum_{k=1}^{n} y_k e^{j\omega k}\right|^2 \, \mathrm{d}\omega \\ &= \sigma^2 \left(\min_{\omega} |C_{\theta_\circ}(e^{j\omega})|^{-2} + \min_{\omega} |D_{\theta_\circ}(e^{j\omega})|^{-2}\right) > 0 \end{aligned}$$

where the last inequality follows since the polynomials involved are continuous (and hence bounded) functions of $\omega$.

Therefore,

$$V_N''(\zeta_N) = P(\theta_\circ) + \Delta_N$$

where $\Delta_N \stackrel{a.s.}{\to} 0$ and $P(\theta_\circ) > 0$ so that for some $N_1 < \infty$, the matrix $V_N''(\zeta_N)$ is invertible for any $N > N_1$ and hence by (19) and (20), for any $\alpha < 1/2$

$$N^\alpha(\theta_\circ - \widehat{\theta}_N) = [V_N''(\zeta_N)]^{-1} N^\alpha V_N'(\theta_\circ) \stackrel{a.s.}{\longrightarrow} 0$$

as $N \to \infty$ so that, analogous to the result of § 4, the convergence rate bound

$$\widehat{\theta}_N - \theta_\circ = o(N^{-\alpha}), \qquad \alpha < 1/2$$

holds with probability one. Again the formulation of Theorem 2.1 has been used to advantage to gain this convergence rate bound.

## 4.3 Stochastic Approximation

Consider the scenario of an observed data sequence $\{y_t\}$ being modelling in a 'linear regression' form

$$y_t = \phi_t^T \theta_\circ + \nu_t$$

where $\theta_\circ \in \mathbf{R}^n$ is a vector of unknown parameters of interest, $\phi_t \in \mathbf{R}^n$ is a vector of measured signals, and $\{\nu_t\}$ is a measurement corruption modelled as a realisation of a stochastic process.

Suppose that estimation of the parameter vector $\theta_\circ$ is attempted via the use of the stochastic approximation algorithm

$$\theta_{t+1} = \theta_t + \frac{1}{t}\phi_t(y_t - \phi_t^T\theta_t). \tag{21}$$

Then simple manipulation of this equation shows that the estimation error $\widetilde{\theta}_t \triangleq \theta_t - \theta_\circ$ obeys, for any $\alpha$, the relationship

$$t^\alpha \widetilde{\theta}_{t+1} = t^\alpha \widetilde{\theta}_t + \frac{1}{t}\left(\phi_t t^\alpha \nu_t - (\phi_t\phi_t^T) t^\alpha \widetilde{\theta}_t\right).$$

Now, if $\{\phi_t\}$ is a (vector) stationary process whose components satisfy the conditions of corollary 3.3 then

$$\frac{1}{N}\sum_{t=1}^{N}\phi_t\phi_t^T \xrightarrow{a.s.} P \triangleq \mathbf{E}\left\{\phi_t\phi_t^T\right\} \qquad \text{as } N\to\infty \tag{22}$$

where the convergence is componentwise in the matrices involved, and it will be assumed that $P > 0$.

As a final set of assumptions, suppose that the stochastic model for $\{\nu_t\}$ is a wide-sense stationary one with associated spectral density $f_\nu(\lambda)$ bounded above by $C' < \infty$ and that $\{\phi_t\}$ is uncorrelated with $\{\nu_t\}$. Then combining this with the previously stated moment bound assumptions on the components of $\phi_t$ leads to

$$\begin{aligned}\sum_{t=1}^{N}\sum_{s=1}^{N} t^\alpha s^\alpha \mathbf{E}\left\{\phi_t(m)\phi_s(m)\right\}\mathbf{E}\left\{\nu_t\nu_s\right\} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f_\nu(\lambda)\mathbf{E}\left\{\left|\sum_{t=1}^{N} t^\alpha\phi_t(m)e^{j\lambda t}\right|^2\right\}\mathrm{d}\lambda \\ &\le C'\sum_{t=1}^{N} t^{2\alpha}\mathbf{E}\left\{\phi_t^2(m)\right\} \le C''N^{1+2\alpha}.\end{aligned}$$

for some $C'' < \infty$. Therefore, by Theorem 2.1, for any $\epsilon > 0$

$$\frac{1}{N^{1/2+\alpha+\epsilon}}\sum_{t=1}^{N} t^\alpha\phi_t(m)\nu_t \xrightarrow{a.s.} 0, \qquad \text{as } N\to\infty. \tag{23}$$

Combining (22) with $\delta = 1/2$, (23) and the use of a recent result of Kouritzin (Lemma A.2) then provides the conclusion that

$$t^{1/2-\epsilon}\widetilde{\theta}_t \xrightarrow{a.s} 0, \qquad \text{as } N\to\infty.$$

for any $\epsilon > 0$. Not only does this establish the strong consistency of the stochastic approximation scheme (21), but it also indicates the rate of convergence as being at least as fast as $1/t^{1/2-\epsilon}$ with probability one.

# 5 Conclusion

The main convergence result of Theorem 2.1, being so straight-forward to establish, and based on results (Lemma A.1) more than thirty years old, it is possible that it may already exist in the probability theory, statistics or time series literature. However, if so, the author is unaware of it, and presumably (given the contemporary frequent reference to the more restrictive Cramér-Leadbetter conditions as sufficient for strong convergence) this also holds true for the wider engineering community that might also find application for theorem 2.1. Hence the paper at hand.

# A Technical Lemmata

**Lemma A.1.** *Let $\{\nu_t\}$ be arbitrary random variables (not necessarily independent or identically distributed). Suppose that there exists a set of non–negative numbers $\{u_k\}$ such that for some $\beta > 1$*

$$\boldsymbol{E}\left\{\left|\sum_{t=n}^{m}\nu_t\right|^2\right\} \leq \left(\sum_{t=n}^{m}u_t\right)^{\beta}.$$

*Then for any $\varepsilon > 0$, there exists a $K < \infty$ that is independent of $n$ such that*

$$\boldsymbol{P}\left\{\max_{1\leq\ell<n}\left|\sum_{t=1}^{\ell}\nu_t\right| \geq \varepsilon\right\} \leq \frac{K}{\varepsilon^2}\left(\sum_{t=1}^{n}u_t\right)^{\beta}.$$

*Proof.* See theorem 12.2 of [2] or Corollary 1 of [21]. □

**Lemma A.2 (Kouritzin).** *Suppose $\{A_t\}$ is a sequence of $n \times n$ matrices for which*

$$\limsup_{N\to\infty}\left\|\frac{1}{N}\sum_{t=1}^{N}A_t - A\right\| = 0$$

*for some (symmetric) positive definite matrix $A$, and suppose that $\{h_t\}$ is an $\mathbf{R}^n$ valued sequence satisfying*

$$h_{t+1} = h_t + \frac{1}{t}(b_t - A_t h_t).$$

*Then a necessary and sufficient condition for $\lim_{t\to\infty} h_t = h$ is that*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}b_t = b \triangleq Ah.$$

*Proof.* See Proposition 1 of [17]. □

# References


[1] R. ASH, *Real Analysis and Probability*, Academic Press, 1972.

[2] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley and Sons., New York, 1968.

[3] L. BREIMAN, *Probability*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1992.

[4] P. CAINES, *Linear Stochastic Systems*, John Wiley and Sons, New York, 1988.

[5] K. CHUNG, *A Course in Probability Theory*, Harcourt, Brace and World Inc., 1968.

[6] H. CRAMÉR AND M. LEADBETTER, *Stationary and Related Stochastic Processes:Sample Function Properties and their applications*, John Wiley and Sons, 1967.

[7] J. DOOB, *Stochastic Processes*, John Wiley and Sons, London, 1953.

[8] R. DUDLEY, *Real Analysis and Probability*, Wadsworth and Brooks/Cole, 1989.

[9] M. EISEN, *Introduction to Mathematical Probabilty Theory*, Prentice Hall, 1969.

[10] E. EWEDA AND O. MACCHI, *Convergence of the RLS and LMS adaptive filters*, IEEE Transactions on Circuits and Systems, 34 (1987), pp. 799–803.

[11] W. A. FULLER, *Introduction to Statistical Time Series*, Probability and Statistics, Wiley-Interscience, New York, 1996.

[12] L. GERENCSÉR, *On a class of mixing processes*, Stochastics, 26 (1989), pp. 165–191.

[13] E. HANNAN, *Multiple Time Series*, John Wiley and Sons, 1970.

[14] E. HANNAN AND M. DEISTLER, *The Statistical Theory of Linear Systems*, John Wiley and Sons, New York, 1988.

[15] A. HELMICKI, C. JACOBSON, AND C. NETT, *Control oriented system identification:A worst case/deterministic approach in $H_\infty$*, IEEE Transactions on Automatic Control, 36 (October 1991), pp. 1163–1176.

[16] A. HEUNIS, *Asymptotic properties of prediction error estimators in approximate system identification*, Stochastics, 24 (1988), pp. 1–43.

[17] M. A. KOURITZIN, *On the convergence of linear stochastic approximation procedures*, IEEE Transactions on Information Theory, 42 (1996), pp. 1305–1309.

[18] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Transactions on Automatic Control, 22 (1977), pp. 551–575.

[19] ——, *System Identification: Theory for the User, (2nd edition)*, Prentice-Hall, Inc., New Jersey, 1999.

[20] L.LJUNG, *Convergence analysis of parametric identification methods*, IEEE Transactions on Automatic Control, AC-23 (1978), pp. 770–783.

[21] M. LONGNECKER AND R. SERFLING, *General moment and probability inequalities for the maximum partial sum*, Acta Mathematica Acadmiae Scientiarum Hungaricae, 30 (1977), pp. 129–133.

[22] F. MÓRICZ, *Moment inequalities and the strong law of large numbers*, Zeitschrift für Wahrscheinlichkeitstheorie and verwandte Gebeite, 35 (1976), pp. 299–314.

[23] P. PHILLIPS AND V. SOLO, *Asymptotics for linear processes*, The Annals of Statistics, 20 (1992), pp. 971–1001.

[24] A. RÉNYI, *Probability Theory*, North Holland Publishing Company, 1970.

[25] P. RÉVÉSZ, *The Laws of Large Numbers*, Academic Press, 1968.

[26] W. F. STOUT, *Almost Sure Convergence*, Academic Press, 1974.

[27] D. W. STROOK, *Probability Theory, An Analytic View*, Cambridge University Press, 1993.

[28] T.SÖDERSTRÖM AND P.STOICA, *System Identification*, Prentice Hall, New York, 1989.

[29] S. VERES AND J. NORTON, *Structure selection for bounded-parameter models: Consistency conditions and selection criterion*, IEEE Transactions on Automatic Control, AC-36 (1991), pp. 474–481.

[30] D. WILLIAMS, *Probability with Martingales*, Cambridge University Press, Cambridge, UK, 1991.